5-1-2016

Vol. 15, No. 1 (Full Issue)

JMASM Editors

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

Recommended Citation

Editors, JMASM (2016) "Vol. 15, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 15 : Iss. 1 , Article 51.

DOI: 10.22237/jmasm/1462078200

Available at: <http://digitalcommons.wayne.edu/jmasm/vol15/iss1/51>

This Full Issue is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized editor of DigitalCommons@WayneState.

```

do i1 = 1,4
  j(1) = i1
  do i2 = 1,4
    j(2) = i2
    do i3 = 1,4
      j(3) = i3
      do i4 = 1,4
        j(4) = i4
        if (j(1) .eq. j(2) .or. j(1) .eq. j(3) .or. j(1) .eq. j(4)) cycle
        if (j(2) .eq. j(3) .or. j(2) .eq. j(4)) cycle
        if (j(3) .eq. j(4)) cycle
        print*,j(1),j(2),j(3),j(4)
      end do
    end do
  end do
end do

```

Journal of Modern Applied Statistical Methods

Invited Articles

Harvey J. Keselman, Joel R. Levin,
Abdul R. Othman,
Ronald C. Serlin, *and* Rand R. Wilcox

Vol. 15, No. 1 • May, 2016

Journal of Modern Applied Statistical Methods

Shlomo S. Sawilowsky

SENIOR EDITOR

College of Education
Wayne State University

Jack Sawilowsky

EDITOR

Reason Statistical Consulting

Harvey Keselman

ASSOCIATE EDITOR EMERITUS

Department of Psychology
University of Manitoba

Alan Klockars

ASSISTANT EDITOR EMERITUS

Educational Psychology
University of Washington

Bruno D. Zumbo

ASSOCIATE EDITOR

Measurement, Evaluation,
& Research Methodology
University of British Columbia

Vance W. Berger

ASSISTANT EDITOR

Biometry Research Group
National Cancer Institute

Todd C. Headrick

ASSISTANT EDITOR

Educational Psychology
& Special Education
So. Illinois University–
Carbondale

Clayton Hayes

EDITORIAL ASSISTANCE

Joshua Neds-Fox

EDITORIAL ASSISTANCE

Heather Marie Perrone

EDITORIAL ASSISTANCE

JMASM (ISSN 1538–9472, <http://digitalcommons.wayne.edu/jmasm>) is an independent, open access electronic journal, published biannually in May and November by JMASM Inc. (PO Box 48023, Oak Park, MI, 48237) in collaboration with the Wayne State University Library System. *JMASM* seeks to publish (1) new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods, (2) the study of nonparametric, robust, permutation, exact, and approximate randomization methods, and (3) applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Journal correspondence (other than manuscript submissions) and requests for advertising may be forwarded to ea@jmasm.com. See back matter for instructions for authors.

Journal of Modern Applied Statistical Methods

Vol. 15, No. 1

✧ May 2016 ✧

Table of Contents

Invited Articles

---

2 – 11	J. R. LEVIN R. C. SERLIN	The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach
12 – 31	R. R. WILCOX	ANCOVA: A Global Test Based On A Robust Measure of Location or Quantiles When There Is Curvature
32 – 61	H. J. KESELMAN A. R. OTHMAN R. R. WILCOX	Generalized Linear Model Analyses For Treatment Group Equality When Data Are Non-Normal
62 – 77	R. R. WILCOX	Comparisons of Two Quantile Regression Smoothers

Regular Articles

---

78 – 99	J. FEYS	New Nonparametric Rank Tests for Interactions In Factorial Designs with Repeated Measures
100 – 122	G. L. BAIRD L. L. HARLOW	Does One Size Fit All? A Case for Context- Driven Null Hypothesis Statistical Testing
123 – 140	C. H. HOLLAND T. HOLLAND	A Two and More Independent Factor Form of the Wilcoxon-Mann-Witney Test, Extendable to Other Permutation-Based Tests

141 – 159	D. T. NGUYEN E. S. KIM P. RODRIGUEZ DE GIL A. KELLERMAN Y. -H. CHEN J. D. KROMREY A. BELLARA	Parametric Tests for Two Population Means under Normal and Non-Normal Distributions
160 – 192	M. DESAI A. A. MITANI S. W. BRYSON T. ROBINSON	Multiple Imputation When Rate of Change is the Outcome of Interest
193 – 205	R. ARUNACHALAM M. SIVASUBRAMANIAN D. K. GHOSH	Construction of Pair wise Balanced Design
206 – 238	B. M. GOLAM KIBRIA S. BANIK	Some Ridge Regression Estimators and Their Performances
239 – 254	R. ARUNACHALAM M. SIVASUBRAMANIAN D. K. GHOSH	Construction of Efficiency-Balanced Design Using Factorial Design
255 – 275	R. D. SZCZESNIAK D. LI R. S. AMIN	Semiparametric Mixed Models for Nested Repeated Measures Applied to Ambulatory Blood Pressure Monitoring Data
276 – 298	L. C. MCCANDLESS M. L. PATTERSON L. B. CURRIE A. MONIRUZZAMAN J. M. SOMERS	Bayesian Estimation of the Size of a Street-Dwelling Homeless Population
299 – 315	N. KOCK	Non-Normality Propagation Among Latent Variables & Indicators in PLS-SEM Simulations
316 – 331	M. AHSANULLAH M. SHAKIL B. M. GOLAM KIBRIA	Characterizations of Continuous Distributions by Truncated Moment
332 – 357	G. L. BAIRD S. L. BIEBER	The Goldilocks Dilemma: Impacts of Multicollinearity—A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models

358 – 372	R. SINGH H. K. VERMA. P. SHARMA	Estimation of Population Mean Using Exponential Type Imputation Technique for Missing Observations
373 – 388	R. LIU D. WU X. ZHANG S. KIM	Compound Identification Using Penalized Linear Regression on Metabolomics
389 – 416	B. LEBEAU	Impact of Serial Correlation Misspecification with the Linear Mixed Model
417 – 451	V. ARYADOUST Z. SHAHSAVAR	Validity of the Persian Blog Attitude Questionnaire: An Evidence-Based Approach
452 – 471	H. LIAO Y. LI G. BROOKS	Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates
472 – 487	C. O. OZGUR	Z and t Distributions in Hypothesis Testing: Unequal Division of Type I Risk
488 – 510	M. A. W. MAHMOUD R. M. EL-SAGHEER A. A. SOLIMAN A. H. ABD ELLAH	Bayesian Estimation of $P[Y < X]$ Based on Record Values from the Lomax Distribution and MCMC Technique
511 – 538	B. F. FRENCH H. FINCH	Factorial Invariance Testing under Different Levels of Partial Loading Invariance within a Multiple Group Confirmatory Factor Analysis Model
539 – 569	J. R. HARRING J. LIU	A Comparison of Estimation Methods for Nonlinear Mixed Effects Models Under Model Misspecification and Data Sparseness: A Simulation Study
570 – 583	G. O. KORTER	A Spatial Analytical Framework For Examining Road Traffic Crashes
584 – 599	S. LIPOVETSKY	Generalized Singular Value Decomposition With Additive Components

600 – 615	R. SINGH S. B. GUPTA S. MALIK	Almost Unbiased Estimator Using Known Value of Population Parameter(s) in Sample Surveys
616 – 634	S. DEVIKA S. GEORGE L. JEYASEELAN	Application of Esscher Transformed Laplace Distribution in Microarray Gene Expression Data
635 – 669	T. A. TARRAY H. P. SINGH	New Procedures of Estimating Proportion and Sensitivity Using Randomized Response in a Dichotomous Finite Population
670 – 689	T. S. KAMBLE D. N. KASHID	Variable Selection in Regression Using Multilayer Feedforward Network
690 – 710	S. ARUMAIRAJAN P. WIJEKOON	Principal Component Preliminary Test Estimator In The Linear Regression Model
711 – 737	O. S. YAYA O. I. SHITTU	Symmetric Variants of Logistic Smooth Transition Autoregressive Models: Monte Carlo Evidences
738 – 751	Y. ASAR	Liu-Type Logistic Estimators With Optimal Shrinkage Parameter
752 – 773	H. DUZAN N. S. B. M. SHARIFF	Solution to the Multicollinearity Problem By Adding Some Constant to the Diagonal
774 – 788	S. SEN S. S. MAITI N. CHANDRA	The xgamma Distribution: Statistical Properties and Application
789 – 801	D. A. A. G. SINGH E. J. LEAVLINE	Model-Based Outlier Detection System With Statistical Preprocessing
802 – 824	E. J. LEAVLINE S. SUTHA	Analysis and Modeling of Statistical Properties Of FMDFB Subband Coefficients
825 – 835	C. A. VALLABADOS S. A. ARUMUGAM	An Evaluation of Pareto, Lognormal and PPS Distributions: The Size Distribution of Cities In Kerala, India

Algorithms and Code

---

836 – 847	J. KORAN T. C. HEADRICK	A Percentile-Based Power Method in SAS: Simulating Multivariate Non-Normal Continuous Distributions
848 – 854	D. A. WALKER T. J. SMITH	Nine Pseudo R^2 Indices For Binary Logistic Regression Models
855 – 867	W. M. AMIR M. SHAFIQ K. MOKHTAR N. A. ALENG H. A. RAHIM Z. ALI	Simple Response Surface Methodology Using RSREG (SAS)
868 – 883	D. A. WALKER	Confidence Intervals for Kendall's Tau With Small Samples
884 – 892	W. M. AMIR M. SHAFIQ H. A. RAHIM P. LIZA A. ALENG Z. ABDULLAH	Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression (SAS)
893 – 913	S. SUNDARAM D. S. PARTHASARATHY	Determination of Optimal Tightened Normal Tightened Plan Using a Genetic Algorithm

Statistical Software Applications & Review

---

914 – 956	Y. LU	Analyzing Different Sampling Designs (SAS)

Emerging Scholars

---

957 – 977	C. G. THOMPSON	Graphing Effects as Fuzzy Numbers In Meta-Analysis

Invited Article

The Composite Hypothesis Contrast Procedure: A Novel Sequential Multiple-Comparison Approach

Joel R. Levin

University of Arizona
Tucson, AZ

Ronald C. Serlin

University of Wisconsin-Madison
Madison, WI

The sequential composite hypothesis contrast multiple-comparison procedure is introduced for comparing two treatment conditions with one or two control conditions on one or two outcome measures. The procedure deserves consideration insofar as its power advantage over other commonly applied multiple-comparison methods can be sizable.

Keywords: Multiple-comparison procedures, sequential hypothesis testing, logical implications, comparison of means, analysis of variance contrasts

---

In the course of a recent research investigation—a single-case intervention study conducted by Hwang, Levin, and Johnson (2016)—we stumbled upon an interesting data-analysis situation that was reminiscent of one that had been considered a generation ago (Levin, Serlin, & Seaman, 1994). To summarize the take-home message of that 1994 article: Starting with a univariate $K = 3$ independent means one-way layout, we demonstrated that: (a) When an initial omnibus hypothesis test (of, for example, “All μ_k are equal”) is rejected based on a Type I error probability of α , (b) if any sub-hypothesis subsumed by the rejected hypothesis is tested at α , then (c) the resulting familywise Type I error probability (α_{FW}) associated with entire set of tested hypotheses is equal to α .

The assertion follows, chronologically, from Fisher’s (1935) least significant difference (LSD) protected multiple-comparison procedure when applied in a three-mean context; Fletcher’s (personal communication, October 3, 1981) perceptive insights about that particular application of the procedure; Shaffer’s (1986) introduction to, and cogent discussion of, the notion of logical implications

---

Dr. Levin is a Professor Emeritus in the Department of Educational Psychology. Email him at: jrlevin@u.arizona.edu. Dr. Serlin is a Professor Emeritus in the Department of Educational Psychology. Email him at rcserlin@wisc.edu.

of subsumed hypotheses; and the Monte Carlo simulation demonstrations of Seaman, Levin, and Serlin (1991), Zhou and Levin (2004), and others.

Consider a snapshot of logical implications in terms of controlling α_{FW} at α through Fisher's two-stage LSD procedure applied to a one-way ANOVA test of the equality of three independent means, μ_A , μ_B , and μ_C . In that situation, there are only three possible configurations of the three population means: (1) all differ from one another; (2) all are equal; and (3) two means are equal but they differ from the third mean. Let us consider each of these possibilities in turn, in the context of performing an omnibus one-way ANOVA F -test based on $\nu_1 = 2$ and $\nu_2 = N - 3$ degrees of freedom.

In Stage 1, the researcher conducts the omnibus F -test of H_0 : All μ_k are equal. If, and only if, that hypothesis is rejected, the researcher proceeds to Stage 2 and applies a t -test to whichever mean differences (i.e., pairwise or complex contrasts) are of interest, each with a Type I error probability of α . If all population means differ, as in (1) above, and the omnibus-test hypothesis is rejected, then no Type I error can be made in the subsequent set of multiple comparisons because a Type I error can occur only when the means being compared are equal. Note that, in theory only, the researcher could declare that all means differ from each other without even conducting formal t -tests of the differences. Similarly, if the omnibus-test hypothesis is not rejected, no Type I error is made because the error incurred would be a Type II error.

If all population means are equal, as in (2) above, then the Stage 1 omnibus F -test provides the required α_{FW} control of the hypothesis tested. If the hypothesis is not rejected, that is a correct decision, no Type I error is committed, and no Stage 2 multiple comparisons are examined. If, on the other hand, the hypothesis is rejected, then a Type I error has been made with probability α . In that case, in Stage 2, any comparisons of interest can subsequently be examined because, with "familywise" referring to "at least one", the Type I error for the family has already been made and so it doesn't matter whether one, two, or a dozen more occur. Note that, in theory only, one could again declare that all means differ from one another without the formal t -tests.

Finally, if only two population means are equal, as in (3) above, if the Stage 1 omnibus hypothesis is not rejected then that again is a Type II error and the process is terminated. If, however, the hypothesis is rejected, then that is a correct decision and no Type I error has been made. Moreover, insofar as there is only one pair of means that are equal, there is only one opportunity for committing a Type I error in the subsequent set of Stage 2 t -tests. Thus, if each test is conducted based on a Type I error of α , then α_{FW} is also equal to α .

COMPOSITE HYPOTHESIS CONTRAST PROCEDURE

After detailing the underlying basis for the Fisher LSD procedure in the one-way ANOVA three-mean case, Levin et al. (1994) provided several extensions to other $\nu_1 = 2$ degree-of-freedom hypothesis-testing situations (e.g., main effects and interactions in 3×2 factorial designs, χ^2 tests in 3×2 contingency tables, Hotelling's T^2 two-group or MANOVA with two dependent variables). It is important to note that the same familywise Type I error control for the Fisher procedure does not hold for $K > 3$ or $\nu_1 > 2$ situations, even though Shaffer's (1986) logical implications and sequential testing procedures do (Levin et al., 1994; Seaman et al., 1991). Subsequently, similar sequential-testing logic associated with Scheffé's (1970) modified multiple-comparison procedure (Klockars & Hancock, 2000) was illustrated and extended by Zhou and Levin (2004) to hypothesis-testing situations with multiple independent or dependent variables (e.g., tests of P partial regression coefficients, K -group MANOVA with P dependent variables).

The Composite Hypothesis Contrast Procedure

In what follows, a novel sequential testing approach is proposed that is fundamentally different from both the Fisher LSD procedure and the planned Bonferroni-type procedures that were comprehensively reviewed by Shaffer (1986), Seaman et al. (1991), and Levin et al. (1994). Yet, the present approach obeys precisely the same type of successive logical implications that was just presented for the Fisher LSD procedure as applied to $\nu_1 = 2$ hypothesis-testing situations. With this new approach, a test of a single degree-of-freedom comparison (what we have termed a “composite hypothesis contrast”) serves as a Stage 1 screening device, which, if proven to be statistically significant, leads directly to a set of logically implied α_{FW} -controlled additional contrasts. The procedure is so named because it essentially provides a framework for testing two linked hypotheses, first in combination and then individually.

The utility of the Stage 1 test of the composite hypothesis contrast is the same as that of initial omnibus tests associated with conventional multiple-comparison procedures, including those of Fisher (1935), Scheffé (1970), and Tukey (1953), among others. Specifically, if the Stage 1 test is statistically significant, it allows for α_{FW} -controlled follow-up testing of two focal hypotheses of interest. The fundamental assumption underlying application of the procedure is that two different experimental conditions are associated with similar differences or effects on the outcome measure(s), relative to other control or comparison conditions. Consider the approach for a few different comparison-of-means situations by beginning with a one-way layout with three independent conditions and a single

dependent variable, as would be applicable for the Fisher LSD procedure that we have been considering. Although the following discussion assumes equal sample-size situations, special comments on unequal sample sizes are included in the final part of the article.

Design 1: Three Conditions, One Outcome Measure.

In the three-condition case with two experimental conditions and one control condition, it is posited that the difference between each experimental condition and the control condition is of a comparable magnitude and in the same direction – but see the addendum that follows. (As an aside, the following discussion could alternatively assume that there is one experimental condition and two control conditions.) In the first stage of the procedure, the two experimental means are combined (i.e., averaged) and tested against the control mean as a composite hypothesis contrast based on a Type I error probability of α , via a t -test with the MS_W based on $\nu = N - K$ serving as an estimator of the population within-group variance. If statistically significant, in the second stage the two experimental conditions' means are separately compared with the mean of the control condition, each based on a Type I error probability of α . It is suggested both the composite hypothesis contrast and the follow-up separate contrasts typically be conducted as one-tailed tests insofar as a researcher would likely not be adopting this procedure without a solid rationale for and understanding of the direction of the treatment effects.

With α_{FW} controlled through logical implications, the procedure affords an efficient alternative to standard procedures for assessing both the aggregated and separate effects of the two experimental conditions. Specifically, the logical implications here are as follows: (1) If, in the population, either of the two experimental means differs from the control mean, then no Type I error is made with the Stage 1 test. Thus, if the Stage 1 hypothesis is rejected, then at most only one Type I error will be made with the two Stage 2 tests. (2) If, in the population, there is no difference between either of the two experimental means and the control mean, then a rejected Stage 1 hypothesis is a Type I error and, following the familywise Type I error concept, it does not matter whether zero, one, or two additional Type I errors occur during the Stage 2 testing.

Addendum. If (1) the outcome measure represents an interval scale, and (2) none of the to-be-described transformed data will fall beyond the measure's attainable upper or lower limits, then predicted experimental vs. control effects in opposite

COMPOSITE HYPOTHESIS CONTRAST PROCEDURE

directions can also be accommodated in the first stage test of the composite hypothesis contrast. For example, suppose it is predicted that the mean of one experimental condition will be higher than the control condition mean ($\mu_{E1} > \mu_C$) and the mean of the other experimental condition will be lower ($\mu_{E2} < \mu_C$). Further suppose that the actual sample means are in the predicted directions, with E1 exceeding C by 10 points and C exceeding E2 by 8 points. In that case, the E2 data could be transformed for the Stage 1 test by adding a constant of 16 (2×8) to all of the scores in that condition. As a result, the E2 mean will now be 8 points above the C mean, rather than 8 points below it, and the E1 and E2 means could be meaningfully combined for the composite hypothesis contrast test in the manner that was just described.

Design 2: Four Conditions, One Outcome Measure.

The composite hypothesis contrast procedure can be applied to test for differences involving four condition means in a manner similar to what was detailed for the three-condition case. Consider, for example, a study with two experimental conditions (E1 and E2) and two control conditions (C1 and C2). In addition, each experimental condition is conceptually linked to its own control condition: (e.g., E1 is linked to C1 and E2 is linked to C2). The researcher is testing for two similar treatment effects, one based on an ultimate comparison of E1 and C1 and the other based on an ultimate comparison of E2 and C2. The omnibus composite hypothesis contrast is initially tested at α in Stage 1 based on a comparison of the average of the E1 and E2 means with the average of the C1 and C2 means. If statistically significant, in Stage 2 the two separate contrasts are each tested at α , with the familywise Type I error rate controlled at α via logical implications analogous to those described for the three-group situation. In that regard, it is important to note that additional comparisons (e.g., of E1 and C2 or of E2 and C1) are not allowed as they would inflate the specified familywise Type I error rate.

Design 3: Two Conditions, Two Outcome Measures.

Now suppose that there are two conditions, experimental and control, and two different outcome measures of interest, X and Y. Moreover, it is assumed that similar treatment effects will be manifested on X and Y. Following the rationale of Marascuilo and Levin (1983) for creating an equally weighted linear combination of separate dependent variables by standardizing and adding (or averaging) them, a researcher could do the same here. In Stage 1, the composite hypothesis contrast procedure would initially compare the experimental and control condition on their

respective mean linear combinations (here, averages) of the X and Y measures, either standardized or unstandardized, depending on how comparable the two measures are assumed to be, based on a Type I error probability of α . If statistically significant, by logical implications in Stage 2 the experimental and control conditions means could be compared on the original X and Y outcome measures separately, each based on α , and thereby controlling α_{FW} at α .

Design 4: Four Conditions, Two Outcome Measures

A situation that incorporates aspects of Designs 2 and 3 was implemented in the previously cited Hwang et al. (2016) study where, in the context of a single-case crossover design (Levin, Ferron, & Gafurov, 2014), four different learning strategies (two experimental and two control) were predicted to have similar effects on two different outcome measures. Moreover, in that single-case design, the outcome measures of interest were the amounts of change/improvement between the baseline (A) phase and the intervention (B) phase of the study. In Stage 1 of the present statistical procedure, based on $\alpha = 0.05$, a one-tailed test of the composite hypothesis contrast proved to be statistically significant ($p = 0.020$). This result indicates that the composite hypothesis contrast (consisting of the two combined experimental strategies vs. the two combined control strategies), as applied to the change on the averaged two outcome measures, represented a detectable effect that was in the predicted direction. In Stage 2, for the two strategies' "comparison of change" tests on the two separate outcome measures, each at $\alpha = 0.05$, although both experimental strategies yielded effects that were in their expected directions, one of these was reasonably large and statistically significant ($p = 0.012$) while the other was considerably smaller and not statistically significant ($p = 0.087$).

The Dangers Lurking Beneath: Power Considerations

Just because the composite hypothesis contrast procedure can be implemented does not indicate that it is statistically advantageous or optimal to do so, relative to alternative α_{FW} -controlled multiple-comparison procedures that could be conducted instead. In particular, statistical power considerations would be advised when determining whether or when to use this approach.

Consider, for instance, the hypothetical examples presented in Table 1. There it is found that with a three-mean effect size defined as $f^2 = \omega^2/(1 - \omega^2)$, both where f is held constant at 0.471 in Parts A and B of Table 1 and as a general rule: (1) when the two averaged experimental means are equal and different from the control

COMPOSITE HYPOTHESIS CONTRAST PROCEDURE

mean (Panel A), Stage 1 of the present composite hypothesis contrast (CHC) approach overpowers at least three of its would-be competitors, namely, Fisher's

Table 1. Stage 1 powers for Fisher's LSD and the composite hypothesis contrast procedure, as well as powers to detect the larger of the two pairwise comparisons for the Holm-Bonferroni and Dunnett Procedures

A. Two means (E1 and E2) equal, each different from the third mean (C) by 1σ ; three-mean effect size given by $f = 0.471$

<i>n</i>	Fisher	Holm (2T)	Holm (1T)	Dunnett (2T)	Dunnett(1T)	CHC (2T)	CHC (1T)
10	0.58	0.46	0.58	0.47	0.60	0.70	0.81
15	0.78	0.66	0.76	0.68	0.78	0.87	0.93
20	0.90	0.80	0.87	0.81	0.88	0.95	0.98

B. All means different in steps of 0.577σ ($E1 > E2 > C$); three-mean effect size given by $f = 0.471$

<i>n</i>	Fisher	Holm (2T)	Holm (1T)	Dunnett (2T)	Dunnett(1T)	CHC (2T)	CHC (1T)
10	0.58	0.59	0.70	0.60	0.72	0.58	0.70
15	0.78	0.80	0.87	0.80	0.88	0.76	0.85
20	0.90	0.91	0.95	0.90	0.95	0.87	0.93

Note: CHC = the present Composite Hypothesis Test; 2T = two-tailed test; 1T = one-tailed test

LSD Stage 1 omnibus test, along with Holm's (1979) sequential Bonferroni procedure and Dunnett's (1955) "each vs. one" multiple-comparison procedure applied to the larger of the two second-stage experimental vs. control comparisons; and (2) when the three means are more equally separated within the three-mean interval (Panel B), the one- and two-tailed test powers of the CHC approach are only slightly lower than those of the corresponding Holm and Dunnett powers, with the CHC approach's one-tailed powers still remaining higher than those of Fisher's LSD test.

In a previous study, Serlin and Mailloux (1999) investigated the analysis of designs with two conditions and two outcome measures, analogous to Design 3 above. They added together the two standardized outcome measures to form a composite that is similar to the composite that was described earlier here. Consistent with the our power results and conclusions, Serlin and Mailloux found that, if the univariate effect sizes associated the two measures are similar, with the smaller being at least half or more in size as the larger, then the Stage 1 screening test on the composite outcome measure followed by univariate tests in Stage 2 (as was presented here) is a more powerful procedure than both the multivariate Hotelling T^2 test and either Holm's (1979) or Shaffer's (1986) "sequentially rejective" procedures. Consequently, we have good reason to believe that, in the

present four-group application described earlier, that if the smaller of the separate E-C comparisons is at least half the size of the larger, the composite hypothesis contrast approach will also be more powerful than the alternative multiple-comparison procedures that were considered here.

Thus, there is a trade-off between the increased power resulting from the composite hypothesis contrast procedure based on the average of two equal or near-equal experimental means and reduced power resulting from a shrunken composite as the two averaged experimental means get further and further apart. In fact, we have determined that, as long as the ratio of the smaller to the larger experimental mean is at least 0.50, then as far as statistical power is concerned the CHC approach would likely be the hypothesis-testing method of choice in this three-mean situation. It is important to note nonetheless that the just-reported powers are not directly comparable. Those associated with the CHC and Fisher's LSD are Stage 1 omnibus test powers and those of Holm and Dunnett are Stage 2 powers for the larger of the two contrasts of interest. Yet it can be concluded that, because the Stage 2 critical values for the CHC are smaller than those for either Holm or Dunnett, if the CHC Stage 1 hypothesis is rejected, then the Stage 2 contrasts will be detected more often with the former procedure than they will with the two latter procedures.

Caveat

When constructing the composite hypothesis contrast, one must exercise caution in calculating the combined group mean in the case of unequal sample sizes, lest one fall prey to confounding due to what is known as Simpson's (1951) paradox. The paradox is perhaps easiest to envision as resulting from a third-variable influence in a two-way layout, wherein the unequal sample sizes are considered a function of a factor not considered in the design. In the earlier discussed Designs 2 and 3 with four conditions and one or two dependent variables, if weighted-by-sample-size means are used to form the Stage 1 composite hypothesis contrast, it is easy to show that, even if the E1 and C1 means were equal, as were the E2 and C2 means, then the combined E and C means in the composite could differ, in which case the logical implications required for the validity of the method do not hold. The solution, of course, would be to create the composite using unweighted means (i.e., the simple average of the E1 and E2 means minus the simple average of the C1 and C2 means).

Conclusion

The two-stage composite hypothesis contrast procedure is not a statistical panacea for all researchers in all multiple-comparison situations. It may, however, represent a useful statistical tool for some researchers in the situations for which it was intended, typically where two experimental treatments are expected to produce comparable effects (relative to one or two control conditions) on one or two outcome measures. The procedure is recommended for those situations because it provides a straightforward, more powerful statistical alternative to other commonly applied multiple-comparison methods.

References

- Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50(272), 1096-1121. doi: [10.1080/01621459.1955.10501294](https://doi.org/10.1080/01621459.1955.10501294)
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65-70. Available from <http://www.jstor.org/stable/4615733>
- Hwang, Y., Levin, J. R., & Johnson, E. W. (2016). Pictorial mnemonic-strategy interventions for children with special needs: Illustration of a multiply randomized single-case crossover design. *Developmental Neurorehabilitation*. Advance online publication. doi: [10.3109/17518423.2015.1100689](https://doi.org/10.3109/17518423.2015.1100689)
- Klockars, A. J., & Hancock, G. R. (2000). Scheffé's more powerful *F*-protected post hoc procedure. *Journal of Educational and Behavioral Statistics*, 25(1), 13-19. doi: [10.3102/10769986025001013](https://doi.org/10.3102/10769986025001013)
- Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods*, 13(2), 2-52. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol13/iss2/2>
- Levin, J. R., Serlin, R. C., & Seaman, M. A. (1994). A controlled, powerful multiple-comparison strategy for several situations. *Psychological Bulletin*, 115(1), 153-159. doi: [10.1037/0033-2909.115.1.153](https://doi.org/10.1037/0033-2909.115.1.153)

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences: A researcher's guide*. Monterey, CA: Brooks/Cole.

Scheffé, H. (1970). Multiple testing versus multiple estimation: Improper confidence sets, estimation of directions and ratios. *Annals of Mathematical Statistics*, 41(1), 1-29. Available from <http://www.jstor.org/stable/2239715>

Seaman, M. A., Levin, J. R., & Serlin, R. C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. *Psychological Bulletin*, 110(3), 577-586. doi: 10.1037/0033-2909.110.3.577

Serlin, R. C., & Mailloux, M. (1999, April). *An empirical comparison of three methods for performing univariate analyses with multivariate data*. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Québec.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395), 826-831. doi: 10.1080/01621459.1986.10478341

Simpson, E. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2), 238-241. Available from <http://www.jstor.org/stable/2984065>

Tukey, J. W. (1953). *The problem of multiple comparisons* (Unpublished manuscript). Princeton University, Princeton, NJ.

Zhou, X., & Levin, J. R. (2004). A note on extending Scheffé's modified multiple comparison procedure to other analysis situations. *Journal of Modern Applied Statistical Methods*, 3(2), 432-442. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol3/iss2/15/>

Invited Article

ANCOVA: A Global Test Based on a Robust Measure of Location or Quantiles When There Is Curvature

Rand R. Wilcox

University of Southern California
Los Angeles, CA

For two independent groups, let $M_j(x)$ be some conditional measure of location for the j^{th} group associated with some variable Y , given that some covariate $X = x$. The goal is to test $H_0: M_1(x) = M_2(x) \quad \forall x \in \{x_1, \dots, x_p\}$ when using a robust or quantile location estimator.

Keywords: ANCOVA, trimmed mean, non-parametric regression, Harrell–Davis estimator, bootstrap methods, comparing quantiles, Well Elderly 2 study

---

Introduction

For two independent groups, consider the situation where, for the j^{th} group ($j = 1, 2$), Y_j is some outcome variable of interest and X_j is some covariate. The classic ANCOVA method assumes that

$$Y_j = \beta_{0j} + \beta_1 X_j + \varepsilon, \quad (1)$$

where β_{0j} and β_1 are unknown parameters and ε is a random variable having a normal distribution with mean zero and unknown variance σ^2 . So the regression lines are assumed to be parallel and the goal is to compare the intercepts based in part on a least squares estimate of the regression lines. It is well known, however, that there are serious concerns with this approach. First, there is a vast literature establishing that methods based on means, including least squares regression, are not robust (e.g., Staudte & Sheather, 1990; Maronna, Martin, & Yohai, 2006; Heritier, Cantoni, Copt, & Victoria-Feser, 2007; Hampel, Ronchetti, Rousseeuw,

---

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at: rwilcox@usc.edu.

& Stahel, 1986; Huber & Ronchetti, 2009; Wilcox, 2012a; 2012b). A general concern is that violations of underlying assumptions can result in relatively poor power and poor control over the Type I error probability. Moreover, even a single outlier can yield a poor fit to the bulk of the points when using least squares regression.

As is evident, one way of dealing with non-normality is to use some rank-based technique. But rank-based ANCOVA methods are aimed at testing the hypothesis of identical distributions (e.g., Lawson, 1983). So when this method rejects, it is reasonable to conclude that the distributions differ in some manner, but the details regarding how they differ, and by how much, are unclear. An alternative way of gaining some understanding of how the groups differ, but certainly not the only way, is to compare the groups using some measure of location. Here the goal is to make inferences about some robust (conditional) measure of location associated with Y .

Yet another fundamental concern with (1) is that the true regression lines are assumed to be straight. Certainly, in some situations, this is a reasonable approximation. When there is curvature, simply meaning that the regression line is not straight, using some obvious parametric regression model might suffice (e.g., include a quadratic term). But this approach can be inadequate, which has led to a substantial collection of nonparametric methods, often called smoothers, for dealing with curvature in a more flexible manner (e.g., Härdle, 1990; Efromovich, 1999; Eubank, 1999; Fox, 2001; Györfi, Kohler, Krzyzk, & Walk, 2002).

Here, the model given by (1) is replaced with the less restrictive model

$$Y_j = M_j(X_j) + \varepsilon_j, \quad (2)$$

where $M_j(X_j)$ is some unknown function that reflects some conditional measure of location associated with Y given that the covariate value is X_j . The random variable ε_j has some unknown distribution with variance σ_j^2 . So, unlike the classic approach where it is assumed that

$$M_j(X_j) = \beta_{0j} + \beta_{1j}X_j,$$

no parametric model for $M_j(X_j)$ is specified and $\sigma_1^2 = \sigma_2^2$ is not assumed. Let $M_j(x)$ be some (conditional) measure of location associated with Y_j given that $X_j = x$. Here, curvature is addressed using a running interval smoother. Roughly, like all

smoothers, the basic strategy is to focus on the X_j values close to x and use the corresponding Y_j values to estimate $M_j(x)$.

An appeal of the running interval smoother is that it is easily applied when using any robust measure of location. The details are given in the next section of this paper. The goal here is to test the global hypothesis

$$H_0 : M_1(x) = M_2(x), \forall x \in \{x_1, \dots, x_p\}, \quad (3)$$

where $x_1, \dots, x_p$ are p values of the covariate chosen empirically in a manner aimed at capturing any curvature that might exist. Roughly, these p values are chosen using a component of the so-called running interval smoother, which is described in the following section. Put in more substantive terms, the goal is to determine whether two groups (e.g., depressive symptoms among males and females) differ when taking into account the possibility that the extent they differ might depend in a non-trivial manner on some covariate (such as the cortisol awakening response).

In the context of ANCOVA, use of the running interval smoother is not new. In particular, Wilcox (1997) proposed and studied a method that tests $H_0: M_1(x_k) = M_2(x_k)$ for each k , $k = 1, \dots, p$. So p hypotheses are tested rather than the global hypothesis corresponding to (3). The method is based in part on Yuen's (1974) method for comparing trimmed means with the familywise error rate (the probability of one or more Type I errors) controlled using a strategy that is similar to Dunnett's (1980) T3 technique. More recently, a bootstrap variation was proposed and studied by Wilcox (2009). Now the familywise error rate can be controlled using some improvement on the Bonferroni method (e.g., Rom, 1990; Hochberg, 1988). The bootstrap method can, in principle, be used with any robust measure of location.

However, a practical concern with testing p individual hypotheses, rather than a global hypothesis, is that power might be relatively low for three general reasons. First, each individual hypothesis uses only a subset of the available data. In contrast, the global hypothesis used here is based on all of the data that are used to test the individual hypotheses. That is, a larger sample size is used suggesting that it might reject in situations where none of individual tests is significant. Second, if for example the familywise error rate is set at 0.05, then Wilcox's method uses a Type I error probability less than 0.05 for the individual tests, which again can reduce power. The third reason has to do with using a confidence region for two or more parameters as opposed to confidence intervals for each individual parameter of interest. It is known that, in various situations, confidence regions can result in a

significant difference even when there are non-significant results for the individual parameters (for an illustration, see for example [Wilcox, 2012b, p. 690](#)). The method proposed here for testing (3) deals with this issue in a manner that is made clear in a later section. Data from the Well Elderly 2 study ([Jackson et al., 2009](#); [Clark et al., 2011](#)) are used to illustrate that the new method can make a practical difference.

Another goal is to include simulation results on comparing (conditional) quartiles. Comparing medians is an obvious way of proceeding. But, in some situations, differences in the tails of two distributions can be more important and informative than comparisons based on a measure of location that is centrally located (e.g., [Doksum & Sievers, 1976](#); [Lombard, 2005](#)). This proved to be the case in the Well Elderly 2 study for reasons explained in a later section.

Note that, rather than testing (3), a seemingly natural goal is to test the hypothesis that $M_1(x) = M_2(x)$ for all possible values of x , not just those values in the set $\{x_1, \dots, x_p\}$. Numerous papers contain results on methods for accomplishing this goal when $M_j(x)$ is taken to be the conditional mean of Y given that $X = x$ ([Wilcox, 2012a, p. 610](#)). But the mean is not robust and evidently little or nothing is known about how best to proceed when using some robust measure of location. [Wilcox \(2012a\)](#) describes a robust method based on a running interval smoother, but the choice for the span (the value of ℓ_j described in the next section) is dictated by the sample size given the goal of controlling the Type I error probability. That is, a suboptimal fit to the data might be needed. The method used here avoids this problem. Here, some consideration was given to an approach where a robust smoother is applied to each group and predicted Y values are computed for all of the observed x values. If the null hypothesis is true, the regression line for the differences $M_1(x) - M_2(x)$, versus x , should have a zero slope and intercept. Several bootstrap methods were considered based on this approach, but control over the Type I error probability was very poor, so no details are provided.

Description of the Proposed Method

Following [Wilcox \(1997\)](#), the general strategy is to approximate the regression lines with a running interval smoother and then use the components of the smoother to test some relevant hypothesis. A portion of the method requires choosing a location estimator. As will be made evident, in principle any robust location estimator could be used, but here attention is focused on only two estimators: a 20% trimmed mean and the quantile estimator derived by [Harrell and Davis \(1982\)](#).

Let $Z_1, \dots, Z_n$ be any n observations. The γ -trimmed mean is

ANCOVA

$$\sum_{i=g+1}^{n-g} Z_{(i)} ,$$

where $Z_{(1)} \leq \dots \leq Z_{(n)}$ are the $Z_{(i)}$ values written in ascending order and g is the greatest integer less than or equal to γn , $0 \leq \gamma < 0.5$. The 20% trimmed mean corresponds to $\gamma = 0.2$. One advantage of the 20% trimmed mean is that its efficiency compares well to the sample mean under normality (e.g., [Rosenberger & Gasko, 1983](#)). But, as we move toward a more heavy-tailed distribution, the standard error of the 20% trimmed mean can be substantially smaller than the standard error of the mean, which can translate into substantially higher power when outliers tend to occur. Another appeal of the 20% trimmed mean over the mean, when testing hypotheses, is that both theory and simulations indicate that the 20% trimmed is better at handling skewed distributions in terms of controlling the Type I error probability. This is not to suggest that the 20% trimmed mean dominates all other robust estimators that might be used. Clearly this is not the case. The only point is that it is a reasonable measure of location to consider for the situation at hand.

The Harrell and Davis ([1982](#)) estimate of the q^{th} quantile uses a weighted average of all the order statistics. Let U be a random variable having a beta distribution with parameters $a = (n + 1)q$ and $b = (n + 1)(1 - q)$, and let

$$v_i = P\left(\frac{i-1}{n} \leq U \leq \frac{i}{n}\right) .$$

The estimate of the q^{th} quantile, based on $Z_1, \dots, Z_n$, is

$$\hat{\theta}_q = \sum v_i Z_{(i)} .$$

In terms of its standard error, Sfakianakis and Verginis ([2006](#)) show that, in some situations, the Harrell-Davis estimator competes well with alternative estimators that again use a weighted average of all the order statistics, but there are exceptions (Sfakianakis and Verginis derived alternative estimators that have advantages over the Harrell–Davis in some situations). But here it was found that, when sampling from heavy-tailed distributions, the standard error of their estimators can be substantially larger than the standard error of $\hat{\theta}_q$.

Comparisons with other quantile estimators are reported by Parrish (1990) and Sheather and Marron (1990), as well as Dielman, Lowry and Pfaffenberger (1994). The only certainty is that no single estimator dominates in terms of efficiency. For example, the Harrell-Davis estimator has a smaller standard error than the usual sample median when sampling from a normal distribution or a distribution that has relatively light tails, but for sufficiently heavy-tailed distributions, the reverse is true (Wilcox, 2012a, p. 87).

To describe the details of the method for testing (3), let (X_{ij}, Y_{ij}) ($i = 1, \dots, n_j$; $j = 1, 2$) be a random sample of size n_j from the j^{th} group. For a chosen value for x , suppose the goal is to estimate $M_j(x)$. Roughly, for each j , compute a measure of location based on the Y_{ij} values for which the corresponding X_{ij} values are close to x . More formally, for fixed j , compute a measure of location based on the Y_{ij} values such that i is an element of the set

$$P_j(x) = \{i : |X_{ij} - x| \leq \ell_j \text{MADN}_j\},$$

where ℓ_j is a constant chosen by the investigator and often called the span, $\text{MADN}_j = \text{MAD}_j / 0.6745$, MAD_j (the median absolute deviation) is the median of $|X_{1j} - m_j|, \dots, |X_{n_jj} - m_j|$, and m_j is the usual sample median based on $X_{1j}, \dots, X_{n_jj}$. Under normality, $\text{MADN}_j = \text{MAD}_j / 0.6745$ estimates the population standard deviation, in which case X_{ij} is close to x if it is within ℓ_j standard deviations from x . Generally, the choice $\ell_j = 0.8$ or 1 gives good results in terms of capturing any curvature, but of course exceptions are encountered. Let $N_j(x)$ be the cardinality of the set $P_j(x)$, and suppose that $M_j(x)$ is estimated with some measure of location based on the Y_{ij} values for which $i \in P_j(x)$. The two regression lines are defined to be comparable at x if, simultaneously, $N_1(x) \geq 12$ and $N_2(x) \geq 12$. The idea is that if the sample sizes used to estimate $M_1(x)$ and $M_2(x)$ are sufficiently large, then a reasonably accurate confidence interval for $M_1(x) - M_2(x)$ can be computed provided a reasonably level robust technique is used. For example, Yuen's (1974) method might be used with a 20% trimmed mean. It is known that, under fairly general conditions, methods for comparing means are not level robust with relatively small sample sizes (see Wilcox, 2012b for details).

For notational convenience, let θ_{jk} be some location estimator based on the Y values for which $i \in P_j(x_k)$. Let $\hat{\delta}_k = \hat{\theta}_{1k} - \hat{\theta}_{2k}$ and let δ_k denote the population analog of $\hat{\delta}_k$ ($k = 1, \dots, p$). Then (3) corresponds to

ANCOVA

$$H_0 : \delta_1 = \dots = \delta_p = 0 \quad (4)$$

The basic strategy for testing (4) is to generate bootstrap samples from each group, compute $\hat{\delta}_k$ based on these bootstrap samples, repeat this B times, and then measure how deeply the null vector 0 is nested in the bootstrap cloud of points via Mahalanobis distance. Based on these distances, results in Liu and Singh (1997) indicate how to compute a p -value.

To elaborate, let (X_{ij}^*, Y_{ij}^*) be a bootstrap sample from the j^{th} group, which is obtained by resampling with replacement n_j pairs of points from (X_{ij}, Y_{ij}) ($i = 1, \dots, n_j, j = 1, 2$). Let $\hat{\delta}_k^*$ be the estimate of δ_k based on the bootstrap samples from the two groups. Repeat this process B times, yielding $\hat{\Delta}_b^* = (\delta_{1b}^*, \dots, \delta_{pb}^*)$, ($b = 1, \dots, B$). Let $\mathbf{S}$ be the covariance matrix based on the B vectors $\hat{\Delta}_b^*$ ($b = 1, \dots, B$). Note that the center of the bootstrap cloud being estimated by these B bootstrap samples is known. It is $\hat{\Delta}$, the estimate of $\Delta = (\delta_1, \dots, \delta_p)$ based on the (X_{ij}, Y_{ij}) values. Let

$$d_b^2 = (\hat{\Delta}_b^* - \hat{\Delta}) S^{-1} (\hat{\Delta}_b^* - \hat{\Delta})'$$

where, for $b = 0$, $\hat{\Delta}_b^*$ is taken to be the null vector $\mathbf{0}$. Then a (generalized) p -value is

$$\frac{1}{B} \sum_{b=1}^B I(d_0^2 \leq d_b^2), \quad (5)$$

where the indicator function $I(d_0^2 \leq d_b^2) = 1$ if $d_0^2 \leq d_b^2$; otherwise $I(d_0^2 \leq d_b^2) = 0$.

There remains the problem of choosing the x_k values. They might be chosen based on substantive grounds, but of course studying this strategy via simulations is difficult at best. Here, we follow Wilcox (1997) and choose $p = 5$ points in a manner suggested by running interval smoother in terms of capturing any curvature in a flexible manner. For notational convenience, assume that, for fixed j , the X_{ij} values are in ascending order. That is, $X_{1j} \leq \dots \leq X_{nj}$. Suppose z_1 is taken to be the smallest X_{i1} value for which the regression lines are comparable. That is, search the first group for the smallest X_{i1} such that $N_1(X_{i1}) \geq 12$. If $N_2(X_{i1}) \geq 12$, in which case the two regression lines are comparable at X_{i1} , and set $x_1 = X_{i1}$. If $N_2(x_1) < 12$,

consider the next largest x_{i1} value and continue until it is simultaneously true that $N_1(X_{i1}) \geq 12$ and $N_2(X_{i1}) \geq 12$. Let i_1 be the smallest integer such that $N_1(x_{i_1}) \geq 12$ and $N_2(x_{i_1}) \geq 12$. Similarly, let x_5 be the largest X_{i1} value for which the regression lines are comparable. That is, x_5 is the largest X_{i1} value such that $N_1(x_{i1}) \geq 12$ and $N_2(x_{i1}) \geq 12$. Let i_5 be the corresponding value of i . Let $i_3 = (i_1 + i_5)/2$, $i_2 = (i_1 + i_3)/2$, and $i_4 = (i_3 + i_5)/2$. Round i_2 , i_3 , and i_4 down to the nearest integer and set $x_2 = X_{i_2}$, $x_3 = X_{i_3}$, and $x_4 = X_{i_4}$.

When the covariate values are chosen in the manner just described, and $p = 5$ separate tests are performed based on some measure of location, this will be called method W henceforth. Computing a p -value using (5), with the goal of performing a global test, will be called method G. Unless stated otherwise, both methods G and W will be based on a 20% trimmed mean.

Note that, in essence, this is a 2-by- p ANOVA design. But for the p levels of the second factor, the groups are not necessarily independent. The reason is that, for any two covariate values, say x_k and x_m , the intersection of the sets $P_j(x_k)$ and $P_j(x_m)$ is not necessarily equal to the empty set. Here, the strategy for dealing with this feature is to model it via a bootstrap method. Another approach would be to divide the data into p independent groups. But there is uncertainty about how this might be done so as to effectively capture any curvature. The approach used here mimics a basic component used by a wide range of smoothers designed to deal with curvature in a flexible manner.

Of course, the obvious decision rule, when using method G, is to reject the null hypothesis if the p -value is less than or equal to the nominal level. When testing at the $\alpha = 0.05$ level, preliminary simulations indicated that this approach performs well in term of controlling the Type I error probability when $p = 3$ and the x_k values are taken to be the quartiles corresponding to the X_{i1} values. But, when $p = 5$ and the x_k values are chosen as just described, the actual level exceeded 0.075 when testing at the $\alpha = 0.05$ level with $n_1 = n_2 = 30$. This problem persisted with $n_1 = n_2 = 50$. However, for the range of distributions considered (described in the following section), the actual level was found to be relatively stable. This suggests using a strategy similar to Gosset's (Student's) approach to comparing means: Assume normality, determine an appropriate critical value using a reasonable test statistic, and continue using this critical value when dealing with non-normal distributions.

Given n_1 and n_2 , this strategy is implemented by first by generating, for each j , n_j pairs of observations from a bivariate normal distribution having a correlation

$\rho = 0$. Based on this generated data, determine $p = 5$ values of the covariate in the manner just described and then compute the p -value given by (5). Denote this p -value by $\hat{p}$. Repeat this process A times, yielding $\hat{p}_1, \dots, \hat{p}_A$. Then an α level critical p -value, say $\hat{p}_c$, is taken to be the α quantile of the $\hat{p}_1, \dots, \hat{p}_A$ values, which here is estimated via the Harrell-Davis estimator. With $A = 1000$ and when a trimmed mean is used, this can be done in 14.8 seconds using an R function, described in the final section of this paper, running on a MacBook Pro. That is, letting p_o denote the p -value based on the observed data, reject (3) if $p_o \leq \hat{p}_c$.

Note that, once $\hat{p}_c$ has been determined, a $1 - \alpha$ confidence region for the vector $\Delta = (\delta_1, \dots, \delta_p)$ can be computed. A confidence region consists of the convex hull containing the $(1 - \hat{p}_c)B \hat{\Delta}_b$ vectors that have the smallest d_b^2 values. As previously indicated, this confidence region provides a perspective on why the global test considered here can have more power than method W. Situations are encountered where the null vector is not contained in the confidence region, yet the confidence intervals for each of the p differences contain zero.

Simulation

Simulations were used to study the small-sample properties of the proposed method with $n_1 = n_2 = 30$. Smaller sample sizes are dubious because this makes it particularly difficult to effectively deal with curvature. Also, finding five covariate values where the groups are comparable can be problematic. That is, $N_j(x)$ might be so small as to make comparisons meaningless. A few results are reported with $n_1 = n_2 = 100$ and 200 as well.

Estimated Type I error probabilities were based on 4000 replications. The estimated critical p -value was based on $A = 1000$ and $B = 500$ bootstrap samples. Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed.

More precisely, the marginal distributions were taken to be one of four g -and- h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. The R function `ghdist`, in Wilcox (2012a), was used to generate observations from a g -and- h distribution. If Z has a standard normal distribution, then by definition

$$V = \frac{\exp(gZ) - 1}{g} \exp(hZ^2/2)$$

($g > 0$) has a g -and- h distribution, where g and h are parameters that determine the first four moments. That is, a g -and- h distribution is a transformation of the standard normal random variable that can be used to generate data having a range of skewness and kurtosis values. If $g = 0$,

$$V = Z \exp(hZ^2/2) .$$

The four distributions used here were the standard normal ($g = h = 0.0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g -and- h distribution are summarized by Hoaglin (1985).

The g -and- h distributions with $h = 0.2$ were chosen in an attempt to span the range of distributions that might be encountered in practice. The idea is that, if method G performs well for what some might regard as an unrealistic departure from normality, this provides some reassurance that it will perform reasonably when dealing with data from an actual study.

Three types of associations were considered. The first two deal with situations where

$$Y_{ij} = \beta X_{ij} + \varepsilon .$$

The two choices for the slope were $\beta = 0$ and 1. The third type of association was $Y = X^2 + \varepsilon$. These three situations are labeled S1, S2, and S3, respectively. The estimated Type I errors were very similar for S1 and S2, so for brevity only the results for S1 are reported.

The X_{ij} values were generated from a standard normal distribution and ε was generated from one of the four g -and- h distributions previously indicated.

Table 1. Some properties of the g -and- h distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

ANCOVA

Table 2. Estimated Type I error probabilities when testing at the $\alpha = 0.05$ level

g	h	Estimator	S1	S3
0.0	0.0	$\gamma = 0.20$	0.048	0.048
0.0	0.0	$q = 0.50$	0.038	0.044
0.0	0.0	$q = 0.75$	0.049	0.048
0.0	0.2	$\gamma = 0.20$	0.022	0.026
0.0	0.2	$q = 0.50$	0.023	0.028
0.0	0.2	$q = 0.75$	0.029	0.028
0.2	0.0	$\gamma = 0.20$	0.040	0.047
0.2	0.0	$q = 0.25$	0.053	0.056
0.2	0.0	$q = 0.50$	0.036	0.044
0.2	0.0	$q = 0.75$	0.046	0.045
0.2	0.2	$\gamma = 0.20$	0.020	0.024
0.2	0.2	$q = 0.25$	0.040	0.040
0.2	0.2	$q = 0.50$	0.022	0.028
0.2	0.2	$q = 0.75$	0.026	0.025

The simulation results are reported in Table 2. As can be seen, when testing at the 0.05 level, the actual level was estimated to be less than or equal to 0.056 among all of the situations considered. Although the seriousness of a Type I error depends on the situation, Bradley (1978) suggests that, as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. Based on this criterion, the only concern is that, for a very heavy-tailed distribution, the estimated level drops below 0.025, the lowest estimate being 0.020. Increasing both sample sizes to 50 corrects this problem. For example, with $g = h = 0.2$ and $\gamma = 0.2$, the estimate for situation S1 increases from 0.020 to 0.034.

Notice that the lowest estimates in Table 2 occur for $\gamma = 0.2$ when $g = h = 0.2$. Simulations were run again with $n_1 = n_2 = 100$ as well as $n_1 = n_2 = 200$ as a partial check on the impact of using larger sample sizes. The estimated Type I error probabilities for these two situations were 0.036 and 0.040, respectively.

As explained, there are at least three reasons to expect that the global test will have more power than method W. The extent to which this is true depends on the situation. To provide at least some perspective, consider the case where the covariate has a normal distribution and the error term has a g -and- h distribution. First consider $g = h = 0$ (normality) and suppose that the first group has $\beta_1 = \beta_0 = 0$ while, for the second group, $Y = 0.5 + \varepsilon$. With $n_1 = n_2 = 50$ and testing at the 0.05 level, the power of method G test was estimated to be 0.51. The probability of rejecting at one or more design points using method W was estimated to be 0.38. If instead $Y = 0.5X + 0.5 + \varepsilon$ for the second group, the power estimates are now 0.75 and 0.66, respectively. If $Y = 0.5X^2 + 0.5 + \varepsilon$, the estimates are 0.89 and 0.78. For

this last situation, if $(g, h) = (0, 0.2)$, the estimates are 0.76 and 0.70. For $(g, h) = (0.2, 0.2)$ the estimates are 0.75 and 0.70. So all indications are that G has more power, with the increase in power estimated to be as high as 0.12 among the situations considered here.

As already noted, a well-known argument for using a 20% trimmed mean, rather than the mean, is that under normality its efficiency compares very well to the mean. However, as we move toward a heavy-tailed distribution, the standard error of the mean can be substantially larger than the standard error of the 20% trimmed. That is, in terms of power, there is little separating the mean and 20% under normality, but for heavier tailed distributions, power might be substantially higher using a 20% trimmed mean. For the situation at hand, consider again $g = h = 0$ and $Y = 0.5 + \varepsilon$, only now method W is applied using means rather than 20% trimmed means. Now power is estimated to 0.43, slightly better than using a 20% trimmed for which power was estimated to be 0.38. Using instead method G, power was estimated to be 0.56. So again, method G offers more power than method W and power is a bit higher compared to using a 20% trimmed mean, which was 0.51. For $(g, h) = (0, 0.2)$, now the power of method W was estimated to 0.25 when using a mean compared to 0.48 when using a 20% trimmed mean. More relevant to the present paper is that, if method G is used with a mean, power is estimated to be 0.30, which is substantially smaller than the estimate of 0.51 when using a 20% trimmed mean.

Illustrations

There is the issue of whether method G can reject when method W does not when dealing with data from an actual study. There is also the issue of whether comparing quartiles makes a practical difference. This section reports results relevant to these issues using data from the Well Elderly 2 study.

A general goal in the Well Elderly 2 study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the impact of intervention on depressive symptoms as measured by the Center for Epidemiologic Studies Depressive Scale (CES-D). The CES-D (Radloff, 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn, Hoberman, & Rosenbaum, 1988; Foley, Reed, Mutran, & DeVellis, 2002). Higher scores indicate a higher level of depressive symptoms. Another dependent variable was the RAND 36-item Health Survey (SF-36), a measure of self-perceived physical health and mental well-being (Hays,

Sherbourne, & Mazel, 1993; McHorney, Ware, & Raozek., 1993). Higher scores reflect greater health and well-being.

Before intervention and six months following intervention, saliva samples were taken at four times over the course of a single day: on rising, 30 min after rising but before taking anything by mouth, before lunch, and before dinner. Then samples were assayed for cortisol. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicate that measures of stress are associated with the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep (i.e. CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake for about an hour). Here, the goal is to compare males and females after intervention based on CES-D and SF-36 measures using the CAR as a covariate.

To illustrate that in practice the global test can reject when method W does not, and that comparing lower or upper quantiles can make a practical difference, consider the goal of comparing males and females based on CES-D measures using CAR as a covariate. No differences are detected based on a 20% trimmed mean or median when using method W as well as the global test proposed here. This remains the case when comparing 0.25 quantiles using a bootstrap version of method W. But when using method G to compare the groups based on the 0.25 quantile, a significant difference is found. That is, there was no significant difference between males and females based on a measure of location intended to reflect the typical response. But the results indicate that there is a sense in which males tend to have even lower CES-D scores than females.

For the SF-36, testing (3) based on the median, a significant difference is found at the 0.05 level. There were 75 males and 171 females after eliminating missing values. Figure 1 shows a plot of the regression lines where the solid lines is the regression line for males. For the males, there were 6 outliers among the CAR values and, for the females, there were 8 outliers (based on a boxplot) which were eliminated from the analysis and are not shown in Figure 1. Eliminating outliers among the independent variable is allowed; it is eliminating outliers among the dependent variable that can cause technical problems. For the situation in Figure 1, a bootstrap version of method W indicates significant differences when CAR is negative (cortisol increases shortly after awakening). In practical terms, the results indicated that the typical male's perceived health and well-being scores are higher among individuals whose cortisol levels increase after awakening. When cortisol decreases, no significant difference between males and females is found. Moreover, there appears to be little or no association between the CAR and SF-36 among

RAND R. WILCOX

women. For men, again there is no significant association when cortisol increases. But when cortisol decreases, a negative association is found. The slope differs significantly from zero, $p = 0.03$, when fitting a straight line regression via a generalization of the Theil-Sen estimator that is designed to handle tied values.

Note that, in [Figure 1](#), there appears to be curvature for the males. A test of the hypothesis that the regression line is straight was performed using the R function `qrchk` in Wilcox (2012b, p. 544). If again the six outliers among the independent variable are eliminated, the hypothesis of a straight line is rejected at the 0.05 level ($p = 0.046$). If the outliers are retained, now $p = 0.005$. So the results suggest that as CAR increases, there is little change in the typical SF-36 value when CAR is negative. But for CAR positive, the typical SF-36 value for males decreases.

Figure 1. Regression lines for predicting perceived health and well-being; the independent variable is the cortisol awakening response and the solid line is the 0.5 quantile regression line for males

ANCOVA

Figure 2. The least squares regression lines for predicting perceived health and well-being using the same data shown in Figure 1; again, the solid line is the regression line for males

To add perspective, Figure 2 shows the least squares regression lines for the same data used in Figure 1. If the classic ANCOVA method is applied, the slopes do not differ significantly at the 0.05 level ($p = 0.16$) and intercepts do differ significantly ($p = 0.008$). But Figure 1 suggests that there is a distinct bend approximately where CAR is equal to -0.1. Indeed, the least squares estimates of the slope for males, based on the CAR values greater -0.1, differs significantly from the slope for females using a method that allows heteroscedasticity, $p = 0.011$. Heteroscedasticity was addressed by estimating the standard errors via the HC4 estimator (see for example Wilcox, 2012a, p. 242). Again, CAR values flagged as outliers by a boxplot were removed. Using instead the Theil-Sen estimator, again the slopes are significantly different, $p = 0.047$.

Conclusion

The indications are that method G avoids Type I errors well above the nominal level. The highest estimated Type I error probability was 0.056 when testing at the 0.05 level. The only known concern is that, when dealing with a very heavy-tailed distribution, the Type I error probability might be less than 0.025 with sample sizes of 30. Increasing the sample sizes to 50, this problem was avoided among the situations considered.

It is unclear under what circumstances some asymptotic result might be used to determine an appropriate critical value. The answer depends on the sample sizes, the span used by the running interval smoother (ℓ_1 and ℓ_2) and the number of covariate values used. But this would seem to be a minor inconvenience in most situations because a critical value can be determined fairly quickly using the method described in the paper. Even with sample sizes of 300, execution time was only 39.5 seconds on a MacBook Pro.

It is not being suggested that method G dominates all approaches relevant to ANCOVA. It seems fairly evident that no single method dominates, one reason being that different methods are sensitive to different features of the data. Rather, method G provides an approach to ANCOVA that might have practical value in various situations, as was illustrated using the Well Elderly data. Here, for example, by dealing with curvature in a flexible manner coupled with a robust measure of location, the results indicated that when CAR is negative, typical SF-36 scores for males tend to be higher than scores for females. The extent to which they differ appears to have little to do with the value of CAR. But, for CAR greater than zero, this is no longer the case. The differences between males and females tend to decrease as CAR increases. Both classic ANCOVA and robust methods indicate that males tend to have higher SF-36 scores. The robust methods provide a more detailed picture regarding when this is the case. Method G is just one tool that helps provide a more detailed understanding of data beyond the non-robust and less flexible approach based on classic ANCOVA methods. Put in broader terms, is there a single number or a single method that tells us everything we would like to know about how groups compare? The answer seems clear: no. Method G is aimed at dealing with this issue.

Finally, R software is available for applying method G. The function `ancGLOB` performs the calculations and is stored on the author's web page. For faster execution time, C++ subroutines have been written that compute the critical p -value. To take advantage of these subroutines, first install the R package `devtools`

ANCOVA

with the R command `install.packages("devtools")`. Then the C++ subroutines can be installed with the following commands:

```
library("devtools")  
install_github( "WRScpp", "mrxiaoh" )
```

It is suggested that these C++ functions be installed using RStudio. Otherwise, installation might fail. Finally, when using the R function `ancGLOB`, set the argument `cpp=TRUE`.

Acknowledgements

The author thanks Xiao He for supplying the C++ code used in this study.

References

- Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)
- Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology*, 80(3), 265-278. doi: [10.1016/j.biopsycho.2008.10.004](https://doi.org/10.1016/j.biopsycho.2008.10.004)
- Clark, F., Jackson, J., Carlson, M., Chou, C., Cherry, B. J., Jordan-Marsh, M.,...Azen, S. P. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, 66(9), 782-790. doi: [10.1136/jech.2009.099754](https://doi.org/10.1136/jech.2009.099754)
- Clow, A., Thorn, L., Evans, P., & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress: The International Journal on the Biology of Stress*, 7(1), 29-37. doi: [10.1080/10253890410001667205](https://doi.org/10.1080/10253890410001667205)
- Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics – Simulation and Computation*, 23(2), 355-371. doi: [10.1080/03610919408813175](https://doi.org/10.1080/03610919408813175)
- Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika*, 63(3), 421-434. doi: [10.1093/biomet/63.3.421](https://doi.org/10.1093/biomet/63.3.421)

Dunnett, C. W. (1980). Pairwise multiple comparisons in the unequal variance case. *Journal of the American Statistical Association*, 75(372), 796-800. doi: [10.1080/01621459.1980.10477552](https://doi.org/10.1080/01621459.1980.10477552)

Efromovich, S. (1999). *Nonparametric curve estimation: Methods, theory and applications*. New York, NY: Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York, NY: Marcel Dekker.

Foley K., Reed P., Mutran, E., & DeVellis, R. F. (2002). Measurement adequacy of the CES-D among a sample of older African-Americans. *Psychiatric Research*, 109(1), 61-69. doi: [10.1016/S0165-1781\(01\)00360-2](https://doi.org/10.1016/S0165-1781(01)00360-2)

Fox, J. (2001). *Multiple and generalized nonparametric regression*. Thousands Oaks, CA: Sage.

Györfi, L., Kohler, M., Krzyzk, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York, NY: Springer Verlag.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York, NY: Wiley.

Härdle, W. (1990). *Applied nonparametric regression*. Econometric Society Monographs (No. 19). Cambridge, UK: Cambridge University Press.

Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3), 635-640. doi: [10.1093/biomet/69.3.635](https://doi.org/10.1093/biomet/69.3.635)

Hays, R. D., Sherbourne, C. D., & Mazel, R. M. (1993). The rand 36-item health survey 1.0. *Health Economics*, 2(3), 217-227. doi: [10.1002/hec.4730020305](https://doi.org/10.1002/hec.4730020305)

Heritier, S., Cantoni, E, Copt, S., & Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. New York, NY: Wiley.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring Data Tables Trends and Shapes* (pp. 461-515). New York, NY: Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800-802. doi: [10.1093/biomet/75.4.800](https://doi.org/10.1093/biomet/75.4.800)

Huber, P. J., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). New York, NY: Wiley.

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S.,...Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: The USC Well Elderly II trial. *Clinical Trials*, 6(1), 90-101. doi: [10.1177/1740774508101191](https://doi.org/10.1177/1740774508101191)

- Lawson, A. (1983). Rank analysis of covariance: Alternative approaches. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 32(3), 331-337. doi: [10.2307/2987939](https://doi.org/10.2307/2987939)
- Lewinsohn, P. M., Hoberman, H. M., & Rosenbaum M. (1988). A prospective study of risk factors for unipolar depression. *Journal of Abnormal Psychology*, 97(3), 251-264. doi: [10.1037/0021-843X.97.3.251](https://doi.org/10.1037/0021-843X.97.3.251)
- Liu, R. G., & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association*, 92(437), 266-277. doi: [10.1080/01621459.1997.10473624](https://doi.org/10.1080/01621459.1997.10473624)
- Lombard, F. (2005). Nonparametric confidence bands for a quantile comparison function. *Technometrics*, 47(3), 364-369. doi: [10.1198/004017005000000184](https://doi.org/10.1198/004017005000000184)
- Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York, NY: Wiley.
- McHorney, C. A., Ware, J. E., & Raozek, A. E. (1993). The MOS 36-item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, 31(3), 247-263. Available from <http://www.jstor.org/stable/3765819>
- Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, 46(1), 247-257. doi: [10.2307/2531649](https://doi.org/10.2307/2531649)
- Radloff, L. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3), 385-401. doi: [10.1177/014662167700100306](https://doi.org/10.1177/014662167700100306)
- Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, 77(3), 663-666. doi: [10.1093/biomet/77.3.663](https://doi.org/10.1093/biomet/77.3.663)
- Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Understanding Robust and Exploratory Data Analysis*. New York, NY: Wiley.
- Sfakianakis, M. E., & Verginis, D. G. (2006). A new family of nonparametric quantile estimators. *Communications in Statistics – Simulation and Computation*, 37(2), 337-345. doi: [10.1080/03610910701790491](https://doi.org/10.1080/03610910701790491)
- Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410), 410-416. doi: [10.1080/01621459.1990.10476214](https://doi.org/10.1080/01621459.1990.10476214)

RAND R. WILCOX

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York, NY: Wiley.

Wilcox, R. R. (1997). ANCOVA based on comparing a robust measure of location at empirically determined design points. *British Journal of Mathematical and Statistical Psychology*, 50(1), 93-103. doi: [10.1111/j.2044-8317.1997.tb01105.x](https://doi.org/10.1111/j.2044-8317.1997.tb01105.x)

Wilcox, R. R. (2009). Robust ANCOVA using a smoother with bootstrap bagging. *British Journal of Mathematical and Statistical Psychology*, 62(2), 427-437. doi: [10.1348/000711008X325300](https://doi.org/10.1348/000711008X325300)

Wilcox, R. R. (2012a). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.

Wilcox, R. R. (2012b). *Modern statistics for the social and behavioral sciences: A practical introduction*. New York, NY: Chapman & Hall/CRC press.

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika*, 61(1), 165-170. doi: [10.1093/biomet/61.1.165](https://doi.org/10.1093/biomet/61.1.165)

Invited Article

Generalized Linear Model Analyses for Treatment Group Equality when Data are Non-Normal

H. J. Keselman

University of Manitoba
Winnipeg, Manitoba

Abdul R. Othman

Universiti Sains Malaysia
Penang, Malaysia

Rand R. Wilcox

University of Southern California
Los Angeles, CA

One of the validity conditions of classical test statistics (e.g., Student's t -test, the ANOVA and MANOVA F -tests) is that data be normally distributed in the populations. When this and/or other derivational assumptions do not hold, the classical test statistic can be prone to too many Type I errors (i.e., falsely rejecting too often) and/or have low power (i.e., failing to reject when the null hypothesis is false) to detect treatment effects when they are present. However, alternative procedures are available for assessing equality of treatment group effects when data are non-normal. For example, researchers can use robust estimators instead of the usual least squares estimators to test that treatment effects are equivalent across groups. As well, recent advances in statistical methodology allow researchers to test for equality of treatment group effects by assuming other distributional shapes for the data. One class of such analyses is generalized linear model techniques. On the other hand, researchers can adopt sequential analyses where they first assess the normality assumption and then depending on the result determine the type of analysis that should be adopted. The purpose of the present study was to compare the above approaches for assessing equality of treatment group effects in the presence of non-normal data. Simulation results which were based on various non-normal distributions and the values of group variances and sample sizes revealed that sequential analysis coupled with a generalized linear model solution were just as prone to inflated or depressed rates of Type I error as the classical ANOVA F -test.

Keywords: Tests for equality of treatment effects, non-normal data, multi-group problem, goodness-of-fit statistic, skewed and kurtotic data, 5-point Likert data, familywise Type I error control

---

Dr. Keselman is Professor Emeritus in the Department of Psychology. Email him at: Harvey.Keselman@umanitoba.ca. Dr. Othman is a Professor of Statistics in the School of Distance Education. Dr. Wilcox is a Professor of Psychology.

Introduction

Classical procedures that test for the equality of treatment effects across independent groups [e.g., Student's Two Independent Sample t-test, the analysis of variance (ANOVA) and multivariate analysis of variance (MANOVA) *F*-tests] provide reliable and valid results when their derivational assumptions hold or nearly hold. All of the aforementioned procedures require that (1) the observations are independent of one another (the independence of observations assumption), (2) the data are distributed normally in each of the treatment populations (the univariate and multivariate normality assumption), and (3) the population variances/covariance matrices are equal across treatment groups (the homogeneity of variances/covariance matrices assumption). When these derivational assumptions hold, the probabilities associated with performing a test of significance are exact; that is, the probabilities of making a Type I error (falsely rejecting the null hypothesis of treatment group equality) and Type II error (falsely accepting the null hypothesis) are known to the researcher. As well, the traditional effect size (ES) statistics and their confidence intervals (CIs) would be valid also.

However, when the assumptions do not hold, either individually or jointly, the actual Type I error probability, as well as the Type II probability, will not be equal to their intended values (e.g., say 0.05 and 0.20, respectively). The consequence is that researchers no longer know the probabilities of inferential error associated with their test of significance for treatment group equality. That is, these errors of inference can be divergently different than those set by the researcher. For example, the probability of committing a Type I error, which is traditionally set at 0.05, could in actuality be 0.60. A consequence of this would be that, when the null hypothesis of treatment group equality has been rejected, researchers would not know whether the hypothesis has been rejected erroneously because the probability of a Type I error is much larger than was set, as a result of violating one or more of the derivational assumptions, or whether the hypothesis was rejected because the population treatment effects are truly not equal across treatment groups. Clearly controlling the probabilities of statistical inference (i.e., Type I and II errors) is paramount when using inferential statistical procedures to assess the credibility of research hypotheses. Moreover the values for ES statistics and their CIs could be inaccurate as well.

Based on the preceding, researchers in the behavioral, biological, and health sciences are encouraged to assess whether derivational assumptions hold prior to using a test of significance (e.g., see [Lix & Keselman, 2009](#); [Lix, Keselman, & Hinds, 2005](#); [Kirk, 2013](#)). Indeed, it is well known that psychological, biological,

and medical data are rarely normal in form (e.g., see Lix & Keselman, 2009; Micceri, 1989). For example, many psychologists collect reaction time data; this data typically will be positively skewed (not symmetric as would be the case with a normal distribution) because of large outlying values in the right hand tail of the distribution indicating large reaction times for some subjects.¹ As well, Likert-type data is frequently collected in social science and medical research and such data is clearly non-normal.

With regard to the assumption that data in the treatment populations are normally distributed, researchers can use a test for normality (e.g., Shapiro-Wilk, Kolmogorov-Smirnov – see D’Agostino & Stephens, 1986; SAS Institute, Inc., 2010b, p. 357). If the results of the preliminary test indicate that the empirical data in each treatment group conforms to a theoretical normal distribution, researchers can go on to test for equality of treatment group effects with the *t*- or *F*-test (assuming that the other assumptions are examined and believed to be true as well); additionally, the size of ES statistics and their CIs would be accurate and meaningful. However, if the result of the test for normality indicates the empirical data are not normally distributed within treatment groups, researchers must take remedial action.

Until recently, the choices available to researchers for dealing with non-normal data were to: (1) transform their data to achieve normality (see Box & Cox, 1964), (2) use a procedure that does not require the data to be normally distributed (a rank transformation test – see Akritas, Arnold, & Brunner, 1997), or (3) use a procedure (Welch-James – WJ_t) that can perform accurately in the presence of non-normality (e.g., procedures that use robust estimators, such as trimmed means and Winsorized variances, rather than the usual least squares estimators in the test statistic) (see Erceg-Hurn, Wilcox, & Keselman, 2013; Keselman, Algina, Lix, Wilcox, & Deering, 2008a; 2008b). Prior empirical research indicates that these alternative methods can be quite successful in controlling Type I and Type II errors, and accurate in calculating ES statistics and setting CIs around them when data are non-normal in the treatment populations (see Keselman et al., 2008a; 2008b; Lix & Keselman, 1998). Nonetheless, some of the better approaches [e.g., alternative (3)] are not available in the major statistical packages and therefore are not typically used by researchers (Keselman et al., 2008b have provided a software program to implement the WJ_t procedure).

However, now researchers have methods for dealing with non-normal distributions when testing hypotheses via generalized linear models (i. e., GLM – see Stroup, 2013). In particular, the SAS Institute, Inc. (2010a; 2010b) and PASW (SPSS, Inc., 2009) systems of statistical programs allows users to conduct tests for

equality of treatment group effects where the data need not be presumed to be normal in form (see e.g. the SAS GLMMIX procedure), through generalized linear models (GLMs) (see Breslow & Clayton, 1993; Nelder & Wedderburn, 1972; Stroup, 2013). Indeed, GLMs allow users to perform tests for treatment group equality for data that can be lognormal, exponential, beta, or gamma, etc. distributed. Furthermore, researchers can even adopt a GLM (e.g., see SASs GLIMMIX procedure) that specifies that data in the treatment groups has varied forms; e.g., in some groups the data are normally distributed while in others they are exponentially distributed. Clearly, this approach for dealing with non-normal data could provide researchers with a remarkably reliable and valid way of testing for treatment group equality where data are not normal in form. However, researchers typically do **not** know how their data are distributed in the parent populations and, accordingly, must rely on (statistical) methods to determine the appropriate shape of their data.² The two most popular statistical packages employed by behavioral sciences and biological researchers (e.g., the SAS and SPSS systems), in addition to providing appropriate plots of the empirical data (e.g., normality probability plots, box-plots, etc.), provide users with test statistics that can be used to test for normality and tests that examine the fit of the data to various theoretical distributions.

Accordingly, whether GLM analyses will work well when adopting this two stage strategy will depend, in part, on how good these preliminary tests for normality and fit perform. Evidence regarding the accuracy of these tests is varied (e.g., see Dufour, Farhat, Gardiol, & Khalaf, 2010; Kowalchuk, Keselman, Wilcox, & Algina, 2006; Keselman, Othman, & Wilcox, 2013; 2014; Rochon & Kieser, 2011; Schoder, Himmelmann, & Wilhelm, 2006). For example, in the one-group pretest-posttest-design, Schoder et al. (2006) found that the Kolmogorov-Smirnov (K-S) goodness of fit test did not have enough power to detect non-normal distributions unless sample size was greater than 100. On the other hand, in the same design, Keselman et al. (2014) found that the Anderson-Darling (A-D) goodness of fit test did have sufficient power (≥ 0.80) if researchers set the level of significance at values greater than 0.05 (e.g., $\alpha = 0.15$ or $\alpha = 0.20$). Again comparing goodness of fit statistics for detecting non-normal distributions in the multi-group (3) problem, Keselman et al. (2014) also found that the A-D test provided good power to detect non-normal distributions even while adopting familywise Type I error control over the three tests of normality with overall significance controlled at 0.15 and 0.20.

Nonetheless, perhaps the most crucial aspect of adopting GLMs is the fact that, in order to use them appropriately (correctly) (e.g., the GLIMMIX procedure),

researchers must correctly specify the ‘link function’ which will determine how the analysis is to be conducted in order to properly assess equality of treatment group effects; a misspecification of the ‘link function’ results in the wrong hypothesis being tested and consequently incorrect results (see [Stroup, 2013, Chapter 3](#)).³

Thus it must be determined whether testing for normality is a good preliminary strategy prior to assessing treatment group equality. It could be the case that uniformly adopting the ANOVA F -test, GLM, or WJ_t might be a better strategy than choosing a method of analysis based on a preliminary test for normality or fit. As well, if users incorrectly specify the link function results will be incorrect ([Cerrito, 2005](#); [Stroup, 2013](#)). That is, we do not know whether this approach to handling non-normal data results in better tests for equality of treatment group effects compared to the previously enumerated methods that are available to researchers. Consequently, the purpose of this investigation is to examine this question.

Methods

A simulation study was conducted to examine GLM analyses when dealing with a single factor one-way univariate design having three groups. Researchers can adopt various GLMs in two ways: One approach first examines the empirical data and compares this data to theoretical distributions, e.g., the normal distribution. Based on this preliminary test, researchers then go on to use GLMs (e.g., GLIMMIX), specifying the shape of the distribution that should be presumed to hold in the population treatment groups (e.g., normal, exponential, lognormal, etc.) through a link function. For example, users could test whether the data are normally distributed in each of the treatment groups and, if the test for normality (SAS provides four tests; users would typically select one – A-D) indicate the data are normally distributed, researchers would then go on to the usual linear model which assumes data are normally distributed (i.e., the ANOVA F -test) when it tests for equality of treatment group effects (SASs general linear model analysis through the General Linear Model Procedure). If, however, the test for normality does not indicate that the data are normally distributed within each treatment group, researchers then frequently try to determine how the data are distributed. They can accomplish this by examining whether the empirical data conform to other theoretical distributions (e.g., lognormal, exponential, gamma, etc.). If, for example, analyses indicates that the data are exponentially distributed by examining fit statistics (e.g., A-D, Cramer-von Mises (CvM), K-S – see [D’Agostino & Stephens, 1986](#); [SAS Institute, Inc., 2010b, p. 279](#)) researchers would then go on to adopt a

GLM model analysis (e.g., GLIMMIX, and through its syntax, tell the program that the data is not normal but exponentially distributed). If one adopts the GLIMMIX procedure it will then perform a test of significance for treatment group equality with estimation procedures that presume the data are exponentially distributed. As indicated in the introduction, GLIMMIX even allows researchers in its estimation methodology to specify that the data are distributed in different forms within each of the treatment groups.

The second manner in which researchers can locate and specify the form of the non-normal data within GLM analyses is to perform a number of analyses specifying different shapes for the data and select the analysis that provides the best fit to the data [SAS allows users to select from a number of fit (Information Criteria) statistics, e.g., Akaike, Schwarz – see Keselman, Algina, Kowalchuk, & Wolfinger, 1998]. That is, within GLM analyses users can request the program to compute fit statistics that then enable users to compare different analyses which specify different shapes for the distribution of the data. Consequently, researchers can compare the fit statistics across the different analyses and pick the results (for the test of equality of treatment group effects) that provided the best fit statistic result.

Thus, GLM model analyses appear to be very versatile tools for examining equality of treatment group effects for data that are not normal in form. However, there is limited published information regarding the effectiveness of the procedure for dealing with non-normal data (see e.g., Schoder et al., 2006; Keselman et al., 2013; 2014); that is, the effectiveness of the procedures for dealing with non-normal data will depend, in part, on how good are the tests for normality, the Information Criteria, and fit in identifying the correct form of the distribution of the data (some published studies report unfavorable results – see e.g., Rochon & Kieser, 2011; Schoder et al., 2006). As a result, if the analyses/procedures cannot identify the correct shape for the distribution(s) of the data to be specified in the link function, GLM analyses will give erroneous conclusions with regard to the test for equality of treatment group effects. However, in many other instances encountered by applied researchers, the nature of the data collected will automatically determine the form of the analysis – the link function. For example, for Likert data, it is obvious that the distribution modelled is the multinomial distribution with cumulative logit as the link function (see SAS Institute, Inc., n.d., p. 64).⁴

Consequently, in our study, we manipulated: (1) the procedure used to assess the shape of a distribution and, in particular, we used the A-D test for normality and also employed the Akaike (1974) fit-statistic (available through the SAS system) (see Keselman, Algina, et al., 1998); (2) the shapes of distributions (15 *g*-and-*h* distributions, 6 contaminated normal mixture models, and 4 multinomial models);

GENERALIZED LINEAR MODEL ANALYSES

(3) the sample sizes: (i) three groups of equal sizes, 20, 40, and 80 per group; (ii) three groups of unequal sizes with total sample size of 120. The g - and h -distributions were simulated for (i) and (ii) (Table 1). However, the contaminated mixed normal and the Likert scale distributions were simulated for total sample size 120 in (i) and (ii) (Tables 2 and 3), and (4) the level of significance for the A-D statistic (i.e., $\alpha = 0.15$ and 0.20).

Choices for non-normal distributions are modifications from Schoder et al. (2006), Zimmerman (2010), and Keselman et al. (2013; 2014). These authors investigated a normal distribution with a single outlier, a normal distribution with 10% of the data containing outliers, skewed distributions with varying skewness, and an ordinal 5-point Likert scale with varying multivariate probabilities (common they state in psychological and medical investigations).

Many non-normal distributions were investigated via g -and- h distributions (see Headrick, Kowalchuk, & Sheng, 2008; Hoaglin, 1983; 1985; Kowalchuk & Headrick, 2010; Tukey, 1960). These distributions with their values for skewness and kurtosis are enumerated in Table 1. We chose a range of values of g and h to cover as broad a spectrum of non-normal distributions that could occur in medical, psychological, and behavioral science experiments (e.g., see Keselman, Huberty, et al., 1998; Micceri, 1989; Wilcox, 2012).

SAS Institute, Inc. (2013) was used to generate g -and- h data. To generate data from a g and h distribution, standard unit normal variables Z_{ij} were converted to g and h distributed random variables via

$$Y_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp\left(\frac{hZ_{ij}^2}{2}\right)$$

when both g and h were non-zero. When g was zero, $Y_{ij} = Z_{ij} \exp\left(\frac{hZ_{ij}^2}{2}\right)$. The Z_{ij}

scores were generated by using RANNOR from SAS (2010a). These equations generate symmetric ($g = 0$) and asymmetric distributions ($g \neq 0$), respectively. As Kowalchuk and Headrick (2010) noted, “The parameter $\pm g$ controls the skew of a distribution in terms of both direction and magnitude. The parameter h controls the tail weight or elongation of a distribution and is positively related with kurtosis” (p. 63). As well, Type I error rates were investigated when data were obtained from a normal distribution [$g = h = 0$, the standard normal distribution (skewness and kurtosis = 0)].

A number of different contaminated mixed-normal distributions, such as those reported in Zimmerman (2010), were examined. Contaminated mixed-normal distributions have one or more outlying values that deviate from the central mean of the distribution by some amount measured in standard deviation units (D). For example, Zimmerman (2010) examined a mixed-normal distribution consisting of samples from $N(0, 1)$ with probability 0.95 and from $N(0, 400)$ with probability 0.05. Tukey (1960) suggested that outliers are a common occurrence in distributions, and others have indicated that skewed distributions frequently depict psychological data (e.g., reaction time data). Accordingly, eight contaminated mixed-normal distributions were examined that had one, two, or four outlying values which were five or ten standard deviations from the mean value. These distributions are enumerated in Table 2.

Finally, like Schoder et al. (2006), and Keselman et al. (2013; 2014), a 5-point Likert scale was simulated; such data is frequently gathered in medical, psychological (e.g., from clinical, personality, and social psychology) and other behavioral science investigations. As well, there has been much discussion in the medical and social science literatures regarding the analysis of outcomes with 5-point Likert data and classical methods of analysis (see e.g., Roberson, Shema, Mundfrom, & Holmes, 1995; Jameson, 2004; Nanna & Sawilowsky, 1998). Consequently, the same conditions as Schoder et al. (2006) and Keselman et al. (2014) were investigated. Specifically, (1) an even distribution ($p = .2$ for each category 1 – 5); (2) a symmetric distribution ($p_1 = 0.1, p_2 = 0.2, p_3 = 0.4, p_4 = 0.2, p_5 = 0.1$); (3) a moderately skewed distribution ($p_1 = 0.5, p_2 = 0.3, p_3 = 0.15, p_4 = 0.04, p_5 = 0.01$); and (4) a heavily skewed distribution ($p_1 = 0.7, p_2 = 0.2, p_3 = 0.06, p_4 = 0.03, p_5 = 0.01$). Thus for the 5-point Likert scale data there were 4 multinomial distributions that were simulated (See Table 3).

We did not investigate the same sample size conditions as Schoder et al. (2006) but did include a reasonable range of values (i.e., $n = 20, 40, 80$) depending on the condition. Specifically,

- (i) For the 5 g -and- h distributions, sample sizes of 20, 40, and 80 were chosen.
- (ii) For 6 contaminated normal distributions, $0.95N(0, 1) + 0.05N(0, k)$, $0.9N(0, 1) + 0.1N(0, k)$, and $0.975N(0, 1) + 0.025N(0, k)$, $k = 25, 100$, a sample size of 40 was chosen.

Because in preliminary testing it would be important to guard against a Type II error (falsely accepting the null hypothesis that the data are normal in form), we

selected significance levels of 0.15, and 0.20 when testing assumptions (we believe that the 0.05 level would not be appropriate in preliminary testing). Because some might find that these levels of significance would be too large with multiple tests for normality (one for each group examined), familywise error control with a level of significance of 0.15 or 0.20 was adopted using Hochberg's (1988) sequentially rejective Bonferroni procedure, setting the overall rate of Type I error at either 0.15 or 0.20. Hochberg's procedure, as indicated, has been found to control the overall rate of Type I error over a set of statistical tests and provides greater power to reject non-null hypotheses than the classical Bonferroni method (see e.g., Hochberg & Tamhane, 1987).

Hochberg's Step-Up Sequentially Acceptive Bonferroni

Hochberg's (1988) step-up Bonferroni procedure is another example of a stepwise Bonferroni method of Type I error control and hence can also be better than the usual Dunn-Bonferroni method (see Kirk, 2013, p. 180). In this procedure, the p -values corresponding to the m statistics for testing the hypotheses $H_{(1)}, \dots, H_{(m)}$ are ordered from smallest to largest $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$. Then, for any $i = m, m-1, \dots, 1$, if $p_i \leq \alpha / (m - i + 1)$, the Hochberg procedure rejects all $H_{i'} (i' \leq i)$. According to this procedure, therefore, one begins by assessing the largest p -value, p_m . If $p_m \leq \alpha$, all hypotheses are rejected. If $p_m > \alpha$, then $H_{(m)}$ is accepted and one proceeds to compare $p_{(m-1)} \leq \alpha / 2$. If $p_{(m-1)} \leq \alpha / 2$, then all $H_i = (i = m-1, \dots, 1)$ are rejected; if not, then $H_{(m-1)}$ is accepted and one proceeds to compare $p_{(m-2)}$ with $\alpha / 3$, and so on.

Each condition in the investigation was replicated 5,000 times. We intended to collect Type I error rates for nine methods of testing treatment group equality (preliminary results indicated that we did not need to collect data for the WJ_t procedure based on trimmed means and Winsorized variances).

Nuances to the Study

Nuances to this study included (1) forcing the estimation and inference method to be maximum likelihood (ML) for all response variables and (2) matching the distributions of the simulated data with those of the response variables, i.e. symmetric data with symmetric response variables and skewed data with the lognormal response variable only (that is, using the proper link function).

Method

When the response variables were modeled using only the “dist=” option in the model statement (from SASs GLIMMIX Procedure), the default estimation and inference method for lognormal and normal responses is restricted maximum likelihood (REML). For the exponential, gamma, inverse Gaussian, multinomial and tcentral (SASs nomenclature for a central t distribution) response variables, the default estimation and inference method is ML. In the present study, only the maximum likelihood method was used; hence ML was invoked with SAS syntax when the response variable was lognormal and normal.

Matching Distribution

Initially, it was observed that the exponential and gamma response variables consistently produced the lowest AIC values, thus dominating the determination of the p -values of these procedures, usually inflating them, at times by an excessive amount, when most of the p -values of GLIMMIX modelled on these responses were less than 0.05. To avoid this anomaly, simulated data were matched with the response variable modelled. Thus, instead of considering the minimum AIC of all modeled link function responses, we considered the minimum AIC of symmetric modeled responses on data simulated from symmetric distributions. In the case of skewed data, we stuck to the lognormal response variable only.

Specific Tcentral Response Variables

The only response variable that can be modeled with parameter values is the tcentral response variable (see [SAS Institute, Inc., n. d., p. 66](#)). For $g = 0$ and $h = 0.225$ simulated data, we decided to model two tcentral responses: $t_{(3)}$ and $t_{(4.039)}$. The subscripted numbers in parentheses represent the degrees of freedom of the tcentral distribution. Three is the default degrees of freedom (in GLIMMIX), while 4.039 was obtained from solving $6/(v - 4) = 154.84$. The left hand side of the equation is the kurtosis of the $t_{(v)}$ distribution while the right hand side is the kurtosis of the $g = 0, h = 0.225$ distribution. Similarly, we also modelled the simulated contaminated mixed-normal distribution with specific tcentral response variables by solving equations involving the variance of the t distribution $6/(v - 4)$ and the variances of the mixed normal distributions. The degrees of freedom are given in [Table 4](#). Note that, for the first two contaminated mixed-normal distributions, the values obtained for the degrees of freedom are less than 0. Hence only $t_{(3)}$ response variables were modelled for them. As for the remaining four, both the $t_{(3)}$ and the t

GENERALIZED LINEAR MODEL ANALYSES

with the calculated degrees of freedom were modelled. Subsequently, all tcentral response variables modelled with calculated degrees of freedom will be known as tcentral2.

Hence the GLIMMIX procedure was employed in the following ways in this study: Specifically, we adapted the GLIMMIX test under the following conditions:

- (1) When the data were modelled on normal responses with the ML method of estimation and inference (this was our baseline measure = the ANOVA F -test.),
- (2) Depending on the result from the A-D test for normality (with $\alpha = 0.15$ and $\alpha = 0.20$), by selecting one of several response variable distributions modelled according to the skew of the simulated data as in Table 1,
- (3) By selecting one of several response variable distributions modelled according to the skew of the simulated data as in Table 1 and depending on the result of the Akaike (1974) model fit statistic,
- (4) Selecting the multinomial option when examining the Likert scale data conditions,
- (5) Selecting the lognormal option with the ML method of estimation and inference when examining data generated from skewed distributions including the g -and- h distributions,⁵
- (6) Depending on the result from the A-D test for lognormality (with $\alpha = 0.15$ and $\alpha = 0.20$), by selecting one of several response variable distributions modelled according to the skew of the simulated data as in Table 1.

Thus, the tcentral2 was modelled in Procedures (2), (3), and (6).

Results

Results were evaluated by adopting Bradley's (1978) criterion for effective Type I error control. According to this perspective, empirical rates of Type I error will be considered well controlled if the empirical value falls within a 0.025 to 0.075 interval. In our tables, values that exceed the upper limit of the interval will be displayed in bold typeface while those values less than the lower limit of the interval will be underscored.

Symmetric Data

Table 5 presents empirical rates of Type I error when the simulated data were symmetric in shape for the three sample sizes investigated ($n_1 = n_2 = n_3 = 20, 40$, or 80). Of the 72 values reported in this table, only four fell outside Bradley's (1978) criterion and all are liberal (> 0.075). The liberal rates ranged in value from 0.0752 to 0.1220. Liberal values occurred when either the A-D normality test was employed or when the Akaike (1974) criterion was used. It should also be noted that rates were well controlled for the baseline test (i.e., the ANOVA F -test) and when the response variable was set at lognormal.

Table 6 presents rates when the simulated data were skewed in shape for the three sample size cases investigated. Interestingly, regardless of the shape of the data, all empirical rates were contained within Bradley's (1978) interval.

Table 7 presents rates of error for non-normal data where the variances and group sizes were unequal across groups and were positively paired with one another. Of the 52 values reported, when unequal variances were in a 1:1:4 ratio (and sample sizes were 30, 40, and 50), only 14 values were not contained in Bradley's (1978) interval. This number increased to 32 when the unequal variances were in a 1:1:16 ratio (sample sizes were 30, 40, and 50). For the same two unequal variances conditions the non-controlled empirical values were 11 and 31, respectively when sample sizes were larger (i.e., 20, 40, and 60).

Table 8 presents rates for the ANOVA F -test and GLIMMIX assuming lognormal data in the link function when sample sizes (30, 40, and 50 or 20, 40, and 60) were paired with unequal variances in either a 1:1:4 or 1:1:16 ratio. Again, one can see from the table that many of the empirical values were either conservative (< 0.025) or liberal (> 0.075). The ANOVA F -test had 12 non-controlled values while GLIMMIX with the lognormal link function had 11 non-controlled rates.

Table 9 presents rates of error for non-normal data when unequal sample sizes (i.e., 30, 40, and 50 or 20, 40, and 60) were paired with unequal variances (4:1:1 or 16:1:1) in a negative fashion. For negative pairings of unequal group sizes and variances, the empirical rates were almost always outside Bradley's (1978) liberal interval and approached values as high as 0.50.

Finally, Table 10 presents empirical rates of Type I error for the ANOVA F -test baseline test and the GLIMMIX test assuming a lognormal link function when unequal variances and group sizes were negatively paired. Of the 28 ANOVA F -test empirical values, only two were contained in Bradley's (1978) interval with liberal values ranging from 0.1046 to 0.9152. Of the 20 GLIMMIX values, all but

one exceeded the upper bound of Bradley's interval with liberal values ranging from 0.1154 to 0.5870.

Conclusion

Generalized linear model analyses (e.g., SASs GLIMMIX procedure) can be used to compare treatment effects across groups when data in the populations are normal or not normal. Having such estimation and testing procedures available to applied researchers should be most beneficial since in applied settings data are not likely to have been drawn from normal populations. Thus, researchers can use generalized linear model analyses to compare treatment effects across groups when data are not normally distributed in the parent populations. This is in contrast to traditional procedures such as the analysis of variance F -test which compares treatment effects across groups but presumes that data are normally distributed in the parent populations. It is well known that the probability of committing a Type I error (falsely rejecting the null hypothesis of treatment group equality) and the probability to detect treatment effects when they are present (the power of the test) are dramatically affected when the normality assumption does not hold. As indicated in the introduction, it is also well known that nonparametric tests are also negatively affected when data are non-normal and variances are unequal (see e.g., [Zimmerman, 2010](#)). Thus, none of these procedures should be adopted to test for equality of central tendency across treatment groups and the availability of alternative generalized analyses is considered an important addition to the researchers arsenal of data analytic techniques.

However, researchers typically do not know how the data in the population distributions are distributed and, accordingly, must rely on informal (e.g., graphs such as normal probability plots) or formal (e.g., tests for normality) methods to determine whether they can adopt the traditional method (i.e., the ANOVA F -test) of investigating for treatment effects across groups, or adopt more modern methods such as a generalized linear model to make such an assessment (i.e., SASs GLIMMIX procedure). Since informal methods are open to subjectivity of the analyzer, formal methods to analyze whether data are normal or not are recommended (e.g., see [SAS Institute, Inc., 2010b](#); [Schoder et al., 2006](#)). Unfortunately, very little is known about the efficacy of adopting formal tests of normality or model fit prior to selecting a method for comparing treatment effects across groups. Prior research regarding the utility of tests for normality are mixed (e.g., see [Dufour et al., 2010](#); [Kowalchuk, et al., 2006](#); [Keselman et al., 2013; 2014](#); [Rochon & Kieser, 2011](#); [Schoder et al., 2006](#)) and the evidence regarding the

efficacy of tests for fit (e.g., Akaike, 1974) are not good (e.g., see Dufour et al., 2010; Keselman, Algina, et al., 2006; Keselman et al., 2013; 2014; Rochon & Kieser, 2011; Schoder et al., 2006). Thus, at this time, we do not know whether adopting this two stage strategy will be effective in testing for treatment group equality.

Accordingly, this problem was investigated by sampling data from various non-normal distributions (skewed distributions having different degrees of skewness and kurtosis, multinomial Likert type distributions, and mixed-normal distributions), either having equal variances or having unequal variances and unequal group sizes that were either positively or negatively paired with one another. The ANOVA F -test was then compared with GLIMMIX results when the GLIMMIX test was adopted following a significant test for non-normality with the Anderson-Darling test statistic or based on the results from applying the Akaike (1974) goodness of fit statistic. A GLIMMIX solution was also adopted in which we always assumed (through the link function) that the data were lognormal in shape.

It is apparent from the empirical findings that adopting a general linear model approach (e.g., GLIMMIX) procedure to assess treatment group equality across groups does not work when the link function of the data is based on the Anderson-Darling test for distribution shape, the smallest Akiake fit-statistic value, or always assuming a log-normal distribution. It is disappointing to report the generalized linear model solutions were not better at controlling the number of Type I errors as compared to the traditional ANOVA F -test. Indeed, for the cases investigated, the ANOVA F -test resulted in an inflated (liberal) or deflated (conservative) rate of Type I error 62 percent of the time while the rate for the GLIMMIX solutions varied from 56 to 61 percent. The lowest percentage of Type I errors for the GLIMMIX solution occurred when we presumed in the analysis that the data were lognormal in the population.

Therefore, researchers should adopt a generalized linear model analysis with caution as it will not necessarily provide better Type I error control when data are non-normal. As was indicated by Cerrito (2005), GLIMMIX is a difficult procedure to adopt, and much thought should be given to choosing this method of analysis. As they stated, “While it is possible to use PROC GLIMMIX as the most complex of the models, it is not advisable. Even so, choices as to random versus fixed effects, link function, and covariance matrix still have to be made. Therefore, the investigator should use the simplest procedure that will accommodate the variable choices.” (p. 7) (see also Stroup, 2013, Chapter 3).

All, however, is not lost. As indicated in the introduction, one can successfully test for equality of central tendency across groups when data are non-normal and variances are heterogeneous by adopting the non-pooled Welch-James statistic (WJ_t) with robust estimators of central tendency and variability. This is a finding that has been established in many research investigations (e.g., see Keselman et al., 2008a; 2008b; Keselman, Wilcox, & Lix, 2003; Keselman, Wilcox, Lix, Algina, & Fradette, 2007; Keselman, Wilcox, Othman, & Fradette, 2002). As well, the WJ_t statistic can be applied with bootstrapping methodology (see Erceg-Hurn et al., 2013; Keselman et al., 2002) resulting in very good Type I error control.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transaction on Automatic Control*, 19(6), 716-723. doi: 10.1109/TAC.1974.1100705
- Akritis, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypotheses and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437), 258-265. doi: 10.1080/01621459.1997.10473623
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2), 211-252. Available from <http://www.jstor.org/stable/2984418>
- Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x
- Breslow, N. R., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421), 9-25. doi: 10.1080/01621459.1993.10594284
- Cerrito, P. B. (2005, October). *From GLM to GLIMMIX-Which model to choose?* Paper presented at the 13th Annual Conference of the Southeast SAS Users Group, Portsmouth, VA.
- D'Agostino, R. B., & Stephens, M. I. A. (Eds.). (1986). *Goodness-of-fit techniques*. New York, NY: Marcel Dekker.
- Dufour, J., Farhat, A., Gardiol, L., & Khalaf, I. (2010). *Simulation-based finite-sample normality tests in linear regressions* (Unpublished paper). Universite de Montreal, Montreal, Québec.

- Erceg-Hurn, D. M., Wilcox, R. R., & Keselman, H. J. (2013). Robust statistical estimation. In T. Little (Ed.), *The Oxford handbook of quantitative methods (Vol. 1)* (pp. 388-406). New York, NY: Oxford University Press.
- Headrick, T. C., Kowalchuk, R. K., & Sheng, Y. (2008). Parametric probability densities and distribution functions for Tukey *g*-and-*h* transformations and their use for fitting data. *Applied Mathematical Sciences*, 2(9), 449-462. Retrieved from <http://www.m-hikari.com/ams/ams-password-2008/ams-password9-12-2008/headrickAMS9-12-2008.pdf>
- Hoaglin, D. C. (1983). *G*-and-*h* distributions. In S. Kotz & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences (Vol. 3)* (pp. 298-301). New York, NY: Wiley
- Hoaglin, D. C. (1985). Summarizing shape numerically: The *g*-and-*h* distribution. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring Data Tables Trends and Shapes* (pp. 461-515). New York, NY: Wiley.
- Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4), 800-803. doi: 10.1093/biomet/75.4.800
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: Wiley.
- Jameson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, 38(12), 1212-1218. doi: 10.1111/j.1365-2929.2004.02012.x
- Keselman, H. J., Algina, J., Kowalchuk, R. K., & Wolfinger, R. D. (1998). Model selection criteria in the analysis of repeated measurements. *The American Statistician*, 60(3), 210-211. doi: 10.1080/03610919808813497
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008a). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, 13(2), 110-129. doi: 10.1037/1082-989X.13.2.110
- Keselman, H. J., Algina, J., Lix, L. M., Wilcox, R. R., & Deering, K. (2008b). *Supplemental material for: A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes*. doi: 10.1037/1082-989X.13.2.110.supp
- Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donahue, B.,...Levin, J. R. (1998). Statistical Practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses. *Educational Research*, 68(3), 350-386. doi: 10.3102/00346543068003350

GENERALIZED LINEAR MODEL ANALYSES

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2013). Preliminary testing for normality: Is it a good practice? *Journal of Modern Applied Statistical Methods*, 12(2), 2-19. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol12/iss2/2/>

Keselman, H. J., Othman, A. R., & Wilcox, R. R. (2014). Testing for normality in the multi-group problem: Is this a good practice? *Clinical Dermatology*, 2(1), 29-43. doi: 10.11138/cderm/2014.2.1.029

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. *Psychophysiology*, 40(4), 586-596. doi: 10.1111/1469-8986.00060

Keselman, H. J., Wilcox, R. R., Lix, L. M., Algina, J., & Fradette, K. (2007). Adaptive robust estimation and testing. *British Journal of Mathematical and Statistical Psychology*, 60(2), 267-293. doi: 10.1348/000711005X63755

Keselman, H. J., Wilcox, R., R., Othman, A. R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality. *Journal of Modern Applied Statistical Methods*, 1(2), 288-309. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol1/iss2/38/>

Kirk, R. E. (2013). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.

Kowalchuk, R., K. & Headrick, T., C. (2010). Simulating multivariate g-and-h distributions. *British Journal of Mathematical and Statistical Psychology*, 63(1), 63-74. doi: 10.1348/000711009X423067

Kowalchuk, R. J., Keselman, H. J., Wilcox, R. R., & Algina, J. (2006). Multiple comparison procedures, trimmed means and transformed statistics. *Journal of Modern Applied Statistical Methods*, 5(1), 44-64. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol5/iss1/6/>

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and non-normality. *Educational and Psychological Measurement*, 58(3), 409-429. doi: 10.1177/0013164498058003004

Lix, L. M., & Keselman, H. J. (2009). Analysis of Variance (ANOVA). In M. W. Kattan (Ed.), *Encyclopedia of medical decision making*. Thousand Oaks, CA: Sage.

- Lix, L. M., Keselman, H. J., & Hinds, A. M. (2005). Robust tests for the multivariate Behrens-Fisher problem. *Computer Methods and Programs in Biomedicine*, 77(2), 129-139. doi: [10.1016/j.cmpb.2004.09.002](https://doi.org/10.1016/j.cmpb.2004.09.002)
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166. doi: [10.1037/0033-2909.105.1.156](https://doi.org/10.1037/0033-2909.105.1.156)
- Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation research. *Psychological Methods*, 3(1), 55-67. doi: [10.1037/1082-989X.3.1.55](https://doi.org/10.1037/1082-989X.3.1.55)
- Nelder, J. A., & Wedderburn, W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3), 370-384. doi: [10.2307/2344614](https://doi.org/10.2307/2344614)
- Roberson, P. K., Shema, S. J., Mundfrom, D. J., & Holmes, T. M. (1995). Analysis of paired Likert data: how to evaluate change and preference questions. *Family Medicine*, 27(10), 671-675. Available from <http://europepmc.org/abstract/med/8582561>
- Rochon, J., & Kieser, M. (2011). A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample *t*-test. *British Journal of Mathematical and Statistical Psychology*, 64(3), 410-426. doi: [10.1348/2044-8317.002003](https://doi.org/10.1348/2044-8317.002003)
- SAS Institute, Inc. (n.d.) *The GLIMMIX Procedure*. In UCLA Statistical Consulting Group. Retrieved from <http://statistics.ats.ucla.edu/stat/sas/glimmix.pdf>.
- SAS Institute, Inc. (2010a). *SAS/STAT 9.22 user's guide*. Cary, NC: SAS Institute, Inc.
- SAS Institute, Inc. (2010b). *Base SAS 9.2 procedures guide: Statistical procedures* (3rd ed). Cary, NC: SAS Institute, Inc.
- SAS Institute, Inc. (2013). *SAS 9.3 for Unix and Linux*. Cary, N C: SAS Institute, Inc.
- Schoder, V., Himmelmann, A., & Wilhelm, K. P. (2006). Preliminary testing for normality: Some statistical aspects of a common concept. *Clinical Dermatology*, 31(6), 757-761. doi: [10.1111/j.1365-2230.2006.02206.x](https://doi.org/10.1111/j.1365-2230.2006.02206.x)
- SPSS, Inc. (2009). *PASW statistics for Windows, version 18.0*. Chicago, IL: SPSS, Inc.
- Stroup, W. W. (2013). *Generalized linear mixed models: Modern concepts, methods and applications*. Boca Raton, FL: CRC Press.

GENERALIZED LINEAR MODEL ANALYSES

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling*. Stanford University Press: Stanford, CA.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.

Zimmerman, D. W. (2010). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. *The Journal of Experimental Education*, 67(1), 55-68. doi: [10.1080/00220979809598344](https://doi.org/10.1080/00220979809598344)

Footnotes

1. It was suggested by a reviewer of an earlier version of this paper that researchers need not concern themselves about normality/non-normality since the Gauss-Markov theorem guarantees that group means are asymptotically normal under much weaker conditions than assuming Gaussian errors. Such a statement surprised us considerably since it is well known that this is not true. Thus, we strenuously maintain that researchers must attend to whether their data are normal or not.
2. It has been suggested by a reviewer of an earlier version of this paper that researchers would know the appropriate link function for their data and would not need to employ tests of fit, etc. to determine the link function in order to get a correct solution. The first and third authors of this paper each have over forty years of experience teaching in departments of psychology and consulting with applied researchers and it is not our experience that researchers would know a priori the correct link function in order to use a generalized linear model analysis correctly.
3. SASs GLIMMIX procedure uses link functions to let the software know what distribution should be fitted to the data. In other words the GLIMMIX procedure assumes that $E[\mathbf{Y} | \boldsymbol{\gamma}] = \mathbf{g}^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma})$, where $\mathbf{g}(\bullet)$ is a differentiable monotonic link function, $\mathbf{Y}$ represents the $(n \times 1)$ vector of observed data, and $\boldsymbol{\gamma}$ is an $(r \times 1)$ vector of random effects (see [SAS Institute, Inc., 2010b, p. 2637](#)). For example, if one were to use the SAS syntax DIST=LOGNORMAL, the GLIMMIX procedure would model the logarithm of the response variable as a normal random variable. Thus, the mean and variance are estimated on the logarithmic scale, assuming a normal distribution, that is, $\log\{\mathbf{Y}\} : N(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ ([SAS Institute, Inc., 2010b, p. 2725](#)).
4. As we indicated in the introduction the link function in GLIMMIX for some types of non-normal data is known. As an example for Likert data the appropriate link function would be the multinomial distribution. For lognormal data the link function would be the identity function. Thus, in many instances applied researchers would know the correct link function (see [SAS Institute, Inc., 2010b](#)).
5. In many psychological investigations reaction time data is collected. Such data is notoriously known to contain large outlying values (i.e., very lengthy delays in reaction times) (see [Tukey, 1960](#)). Accordingly,

GENERALIZED LINEAR MODEL ANALYSES

for such data it would be reasonable to assume that the underlying distribution is lognormal and accordingly we want to investigate the outcome of always assuming that the underlying distribution is of this form when adopting GLIMMIX.

Appendix: Tables

Table 1. Distributions used in this study with their corresponding measures of skewness and kurtosis values

Distribution	Skewness	Kurtosis
N(0, 1)	0.00	0.00
$g = 0, h = 0.225$	0.00	154.84
$g = 0.5, h = 0$ (lognormal)	1.75	8.90
$g = 1, h = 0$ (lognormal)	6.19	110.94
$g = 0.5, h = 0.5$	120.10 ^a	18393.60 ^a

Note: ^aUndefined skewness and kurtosis calculated from 100,000 simulated values

Table 2. Contaminated mixed-normal distributions

(n_1, n_2, n_3)	Group Variances	Distribution	Outliers	
			D	#
(40, 40, 40)	(1, 1, 1)	(0.975)N(0, 1) + (0.025)N(0, 25)	5	(1, 1, 1)
(40, 40, 40)	(1, 1, 1)	(0.950)N(0, 1) + (0.050)N(0, 25)	5	(2, 2, 2)
(40, 40, 40)	(1, 1, 1)	(0.900)N(0, 1) + (0.100)N(0, 25)	5	(4, 4, 4)
(40, 40, 40)	(1, 1, 1)	(0.975)N(0, 1) + (0.025)N(0, 100)	10	(1, 1, 1)
(40, 40, 40)	(1, 1, 1)	(0.950)N(0, 1) + (0.050)N(0, 100)	10	(2, 2, 2)
(40, 40, 40)	(1, 1, 1)	(0.900)N(0, 1) + (0.100)N(0, 100)	10	(4, 4, 4)
(30, 40, 50)	(1, 1, 4)	(0.975)N(0, 1) + (0.025)N(0, 25)	5	(1, 1, 1)
(30, 40, 50)	(1, 1, 4)	(0.950)N(0, 1) + (0.050)N(0, 25)	5	(2, 2, 3)
(30, 40, 50)	(1, 1, 4)	(0.900)N(0, 1) + (0.100)N(0, 25)	5	(3, 4, 5)
(30, 40, 50)	(1, 1, 4)	(0.975)N(0, 1) + (0.025)N(0, 100)	10	(1, 1, 1)
(30, 40, 50)	(1, 1, 4)	(0.950)N(0, 1) + (0.050)N(0, 100)	10	(2, 2, 3)
(30, 40, 50)	(1, 1, 4)	(0.900)N(0, 1) + (0.100)N(0, 100)	10	(3, 4, 5)
(30, 40, 50)	(1, 1, 16)	(0.975)N(0, 1) + (0.025)N(0, 25)	5	(1, 1, 2)
(30, 40, 50)	(1, 1, 16)	(0.950)N(0, 1) + (0.050)N(0, 25)	5	(1, 2, 3)
(30, 40, 50)	(1, 1, 16)	(0.900)N(0, 1) + (0.100)N(0, 25)	5	(2, 4, 6)
(30, 40, 50)	(1, 1, 16)	(0.975)N(0, 1) + (0.025)N(0, 100)	10	(1, 1, 2)
(30, 40, 50)	(1, 1, 16)	(0.950)N(0, 1) + (0.050)N(0, 100)	10	(1, 2, 3)
(30, 40, 50)	(1, 1, 16)	(0.900)N(0, 1) + (0.100)N(0, 100)	10	(2, 4, 6)
(20, 40, 60)	(1, 1, 4)	(0.975)N(0, 1) + (0.025)N(0, 25)	5	(1, 1, 1)
(20, 40, 60)	(1, 1, 4)	(0.950)N(0, 1) + (0.050)N(0, 25)	5	(2, 2, 3)
(20, 40, 60)	(1, 1, 4)	(0.900)N(0, 1) + (0.100)N(0, 25)	5	(3, 4, 5)
(20, 40, 60)	(1, 1, 4)	(0.975)N(0, 1) + (0.025)N(0, 100)	10	(1, 1, 1)
(20, 40, 60)	(1, 1, 4)	(0.950)N(0, 1) + (0.050)N(0, 100)	10	(2, 2, 3)
(20, 40, 60)	(1, 1, 4)	(0.900)N(0, 1) + (0.100)N(0, 100)	10	(3, 4, 5)
(20, 40, 60)	(1, 1, 16)	(0.975)N(0, 1) + (0.025)N(0, 25)	5	(1, 1, 2)
(20, 40, 60)	(1, 1, 16)	(0.950)N(0, 1) + (0.050)N(0, 25)	5	(1, 2, 3)
(20, 40, 60)	(1, 1, 16)	(0.900)N(0, 1) + (0.100)N(0, 25)	5	(2, 4, 6)
(20, 40, 60)	(1, 1, 16)	(0.975)N(0, 1) + (0.025)N(0, 100)	10	(1, 1, 2)
(20, 40, 60)	(1, 1, 16)	(0.950)N(0, 1) + (0.050)N(0, 100)	10	(1, 2, 3)

GENERALIZED LINEAR MODEL ANALYSES

Table 2, continued.

(n_1, n_2, n_3)	Group Variances	Distribution	Outliers	
			D	#
(20, 40, 60)	(1, 1, 16)	$(0.900)N(0, 1) + (0.100)N(0, 100)$	10	(2, 4, 6)
(30, 40, 50)	(4, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 25)$	5	(1, 1, 1)
(30, 40, 50)	(4, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 25)$	5	(2, 2, 3)
(30, 40, 50)	(4, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 25)$	5	(3, 4, 5)
(30, 40, 50)	(4, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 100)$	10	(1, 1, 1)
(30, 40, 50)	(4, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 100)$	10	(2, 2, 3)
(30, 40, 50)	(4, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 100)$	10	(3, 4, 5)
(30, 40, 50)	(16, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 25)$	5	(1, 1, 2)
(30, 40, 50)	(16, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 25)$	5	(1, 2, 3)
(30, 40, 50)	(16, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 25)$	5	(2, 4, 6)
(30, 40, 50)	(16, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 100)$	10	(1, 1, 2)
(30, 40, 50)	(16, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 100)$	10	(1, 2, 3)
(30, 40, 50)	(16, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 100)$	10	(2, 4, 6)
(20, 40, 60)	(4, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 25)$	5	(1, 1, 1)
(20, 40, 60)	(4, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 25)$	5	(2, 2, 3)
(20, 40, 60)	(4, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 25)$	5	(3, 4, 5)
(20, 40, 60)	(4, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 100)$	10	(1, 1, 1)
(20, 40, 60)	(4, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 100)$	10	(2, 2, 3)
(20, 40, 60)	(4, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 100)$	10	(3, 4, 5)
(20, 40, 60)	(16, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 25)$	5	(1, 1, 2)
(20, 40, 60)	(16, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 25)$	5	(1, 2, 3)
(20, 40, 60)	(16, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 25)$	5	(2, 4, 6)
(20, 40, 60)	(16, 1, 1)	$(0.975)N(0, 1) + (0.025)N(0, 100)$	10	(1, 1, 2)
(20, 40, 60)	(16, 1, 1)	$(0.950)N(0, 1) + (0.050)N(0, 100)$	10	(1, 2, 3)
(20, 40, 60)	(16, 1, 1)	$(0.900)N(0, 1) + (0.100)N(0, 100)$	10	(2, 4, 6)

Note: n_i stands for the number of observations in group i ; D stands for distance in standard deviation units; # stands for the number of outliers in the respective groups

Table 3. Multinomial distributions based upon Schoder et al.'s (2006) probabilities simulated as Likert scales

(n_1, n_2, n_3)	Group Variances	Description	$(p_1, p_2, p_3, p_4, p_5)$
(40, 40, 40)	(1, 1, 1)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(40, 40, 40)	(1, 1, 1)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(40, 40, 40)	(1, 1, 1)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(40, 40, 40)	(1, 1, 1)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(30, 40, 50)	(1, 1, 4)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(30, 40, 50)	(1, 1, 4)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(30, 40, 50)	(1, 1, 4)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(30, 40, 50)	(1, 1, 4)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(30, 40, 50)	(1, 1, 16)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(30, 40, 50)	(1, 1, 16)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)

Table 3, continued.

(n_1, n_2, n_3)	Group Variances	Description	$(p_1, p_2, p_3, p_4, p_5)$
(30, 40, 50)	(1, 1, 16)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(30, 40, 50)	(1, 1, 16)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(20, 40, 60)	(1, 1, 4)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(20, 40, 60)	(1, 1, 4)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(20, 40, 60)	(1, 1, 4)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(20, 40, 60)	(1, 1, 4)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(20, 40, 60)	(1, 1, 16)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(20, 40, 60)	(1, 1, 16)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(20, 40, 60)	(1, 1, 16)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(20, 40, 60)	(1, 1, 16)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(30, 40, 50)	(4, 1, 1)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(30, 40, 50)	(4, 1, 1)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(30, 40, 50)	(4, 1, 1)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(30, 40, 50)	(4, 1, 1)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(30, 40, 50)	(16, 1, 1)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(30, 40, 50)	(16, 1, 1)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(30, 40, 50)	(16, 1, 1)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(30, 40, 50)	(16, 1, 1)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(20, 40, 60)	(4, 1, 1)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(20, 40, 60)	(4, 1, 1)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(20, 40, 60)	(4, 1, 1)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(20, 40, 60)	(4, 1, 1)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)
(20, 40, 60)	(16, 1, 1)	Even	(0.20, 0.20, 0.20, 0.20, 0.20)
(20, 40, 60)	(16, 1, 1)	Symmetric	(0.10, 0.20, 0.40, 0.20, 0.10)
(20, 40, 60)	(16, 1, 1)	Moderate skew	(0.50, 0.30, 0.15, 0.04, 0.01)
(20, 40, 60)	(16, 1, 1)	Heavy skew	(0.70, 0.20, 0.06, 0.03, 0.01)

Note: See the note from Table 2

Table 4. Calculated degrees of freedom for the tcentral response variables from the variances of the simulated contaminated mixed normal distributions

Distribution	Variance	Calculated df ^a
(0.975)N(0, 1) + (0.025)N(0, 25)	0.9663	_ ^b
(0.950)N(0, 1) + (0.050)N(0, 25)	0.9650	_ ^b
(0.900)N(0, 1) + (0.100)N(0, 25)	1.0600	35.333
(0.975)N(0, 1) + (0.025)N(0, 100)	1.0131	154.323
(0.950)N(0, 1) + (0.050)N(0, 100)	1.1525	15.115
(0.900)N(0, 1) + (0.100)N(0, 100)	1.8100	4.469

Note: ^adf = degree of freedom; ^bNegative values, only the default was used

GENERALIZED LINEAR MODEL ANALYSES

Table 5. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method=ML) and various versions of GLIMMIX on symmetric data

Sample Size	Distribution	GLIMMIX				
		Baseline	A-D Normal Test		Akaike	Lognormal
			$\alpha = 0.15$	$\alpha = 0.20$		
20	$g = 0.000, h = 0.000$	0.0598	0.0714	0.0726	0.0682	0.0590
	$g = 0.000, h = 0.225$	0.0544	0.0582	0.0574	0.0616	0.0540
40	$g = 0.000, h = 0.000$	0.0532	0.0636	0.0670	0.0572	0.0530
	$g = 0.000, h = 0.225$	0.0488	0.0554	0.0550	0.0558	0.0500
	(0.975)N(0, 1) + (0.025)N(0, 25)	0.0470	0.0744	0.0752	0.0742	0.0466
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.0472	0.0698	0.0696	0.0692	0.0460
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.0484	0.0600	0.0604	0.0604	0.0484
	(0.975)N(0, 1) + (0.025)N(0, 100)	0.0276	0.0736	0.0736	0.0756	0.0266
	(0.950)N(0, 1) + (0.050)N(0, 100)	0.0274	0.0622	0.0622	0.0622	0.0270
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.0416	0.0442	0.0442	0.0442	0.0370
	Likert-Even	0.0520	0.1220	0.1220	0.0520	0.0514
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.0460 ^a				
	Likert-Symmetric	0.0518	0.0550	0.0550	0.0522	0.0532
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.0448 ^a				
80	$g = 0.000, h = 0.000$	0.0572	0.0676	0.0694	0.0574	0.0560
	$g = 0.000, h = 0.225$	0.0520	0.0530	0.0532	0.0532	0.0522

Note: ^aGLIMMIX modelled by multinomial response variable

Table 6. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method = ML) and GLIMMIX with lognormal response on skewed data

Sample Size	Distribution	GLIMMIX	
		Baseline	Lognormal
20	$g = 0.5, h = 0.0$	0.0528	0.0540
	$g = 1.0, h = 0.0$	0.0430	0.0434
	$g = 0.5, h = 0.5$	0.0364	0.0370
40	$g = 0.5, h = 0.0$	0.0544	0.0552
	$g = 1.0, h = 0.0$	0.0442	0.0462
	$g = 0.5, h = 0.5$	0.0330	0.0340
	Likert-Moderate Skew	0.0482	0.0520
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.0442 ^a	
	Likert-Heavy Skew	0.0502	0.0518
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.0392 ^a	
80	$g = 0.5, h = 0.0$	0.0540	0.0534
	$g = 1.0, h = 0.0$	0.0466	0.0498
	$g = 0.5, h = 0.5$	0.0320	0.0336

Note: ^aGLIMMIX modelled by multinomial response variable

Table 7. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method = ML) and various versions of GLIMMIX on symmetric data when group sample sizes are positively paired with group variances

Sample Size/Variance	Distribution	GLIMMIX				
		Baseline	A-D Normal Test		Akaike	Lognormal
			$\alpha=0.15$	$\alpha=0.20$		
(30, 40, 50) / (1, 1, 4)	$g = 0.000, h = 0.000$	0.0408	0.0568	0.0582	0.0762	0.0442
	$g = 0.000, h = 0.225$	0.0368	0.0536	0.0540	0.0554	0.0384
	(0.975)N(0, 1) + (0.025)N(0, 25)	0.0336	0.0742	0.0748	0.0792	0.0330
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.0316	0.0712	0.0716	0.0736	0.0298
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.0304	0.0570	0.0570	0.0572	0.0286
	(0.975)N(0, 1) + (0.025)N(0, 100)	<u>0.0232</u>	0.0772	0.0772	0.0782	<u>0.0236</u>
	(0.950)N(0, 1) + (0.050)N(0, 100)	<u>0.0198</u>	0.0622	0.0622	0.0622	<u>0.0196</u>
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.0270	0.0420	0.0420	0.0420	<u>0.0226</u>
	Likert-Even	0.0382	0.1444	0.1444	0.0382	0.1358
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.0548 ^a				
	Likert-Symmetric	0.0406	0.0552	0.0552	0.0550	0.1380
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.0498 ^a				
(30, 40, 50) / (1, 1, 16)	$g = 0.000, h = 0.000$	0.0424	0.0722	0.0774	0.1462	0.0720
	$g = 0.000, h = 0.225$	0.0396	0.0910	0.0912	0.0938	0.0490
	(0.975)N(0, 1) + (0.025)N(0, 25)	0.0346	0.1176	0.1184	0.1346	0.0392
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.0290	0.1166	0.1168	0.1198	0.0318
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.0316	0.1004	0.1004	0.1008	0.0290
	(0.975)N(0, 1) + (0.025)N(0, 100)	<u>0.0246</u>	0.1276	0.1276	0.1302	0.0308
	(0.950)N(0, 1) + (0.050)N(0, 100)	<u>0.0182</u>	0.1044	0.1044	0.1048	<u>0.0218</u>
	(0.900)N(0, 1) + (0.100)N(0, 100)	<u>0.0236</u>	0.0798	0.0798	0.0796	0.0304
	Likert-Even	0.0424	0.2188	0.2188	0.2186	<u>0.2402</u>
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.0814^a				
	Likert-Symmetric	0.0440	0.0292	0.0292	0.0292	0.0628
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.0576 ^a				
(20, 40, 60) / (1, 1, 4)	$g = 0.000, h = 0.000$	0.0258	0.0364	0.0378	0.0564	0.0278
	$g = 0.000, h = 0.225$	<u>0.0252</u>	0.0400	0.0400	0.0402	0.0254
	(0.975)N(0, 1) + (0.025)N(0, 25)	<u>0.0218</u>	0.0544	0.0550	0.0578	0.0216
	(0.950)N(0, 1) + (0.050)N(0, 25)	<u>0.0158</u>	0.0514	0.0512	0.0520	0.0162
	(0.900)N(0, 1) + (0.100)N(0, 25)	<u>0.0186</u>	0.0390	0.0394	0.0396	0.0180
	(0.975)N(0, 1) + (0.025)N(0, 100)	<u>0.0170</u>	0.0532	0.0532	0.0536	0.0180
	(0.950)N(0, 1) + (0.050)N(0, 100)	<u>0.0102</u>	0.0486	0.0486	0.0486	0.0108
	(0.900)N(0, 1) + (0.100)N(0, 100)	<u>0.0148</u>	0.0290	0.0290	0.0290	0.0154
	Likert-Even	<u>0.0216</u>	0.1084	0.1084	0.0216	0.1046
	(0.2, 0.2, 0.2, 0.2, 0.2)	<u>0.0322^a</u>				
	Likert-Symmetric	<u>0.0228</u>	0.0418	0.0418	0.0338	0.0960
	(0.1, 0.2, 0.4, 0.2, 0.1)	<u>0.0328^a</u>				

GENERALIZED LINEAR MODEL ANALYSES

Table 7, continued.

Sample Size/Variance	Distribution	GLIMMIX				
		Baseline	A-D Normal Test		Akaike	Lognormal
			$\alpha=0.15$	$\alpha=0.20$		
(20, 40, 60) / (1, 1, 16)	$g = 0.000, h = 0.000$	<u>0.0202</u>	0.0384	0.0418	0.0954	0.0390
	$g = 0.000, h = 0.225$	<u>0.0158</u>	0.0428	0.0432	0.0440	<u>0.0186</u>
	(0.975)N(0, 1) + (0.025)N(0, 25)	<u>0.0138</u>	0.0762	0.0774	0.0864	<u>0.0166</u>
	(0.950)N(0, 1) + (0.050)N(0, 25)	<u>0.0122</u>	0.0774	0.0780	0.0804	<u>0.0138</u>
	(0.900)N(0, 1) + (0.100)N(0, 25)	<u>0.0136</u>	0.0614	0.0618	0.0620	<u>0.0112</u>
	(0.975)N(0, 1) + (0.025)N(0, 100)	<u>0.0096</u>	0.0788	0.0792	0.0802	<u>0.0104</u>
	(0.950)N(0, 1) + (0.050)N(0, 100)	<u>0.0052</u>	0.0688	0.0688	0.0688	<u>0.0088</u>
	(0.900)N(0, 1) + (0.100)N(0, 100)	<u>0.0068</u>	0.0444	0.0444	0.0438	<u>0.0118</u>
	Likert-Even	<u>0.0164</u>	0.1614	0.1614	0.1358	0.1900
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.0466 ^a				
	Likert-Symmetric	<u>0.0198</u>	0.0272	0.0272	0.0272	0.0342
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.0314 ^a				

Note: ^aGLIMMIX modelled by multinomial response variable

Table 8. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method=ML) and GLIMMIX with lognormal response on skewed data when group sample sizes are positively paired with group variances

Sample Sizes/Variances	Distribution	GLIMMIX	
		Baseline	Lognormal
(30, 40, 50) / (1, 1, 4)	$g = 0.5, h = 0.0$	0.0416	0.0484
	$g = 1.0, h = 0.0$	0.0552	0.0642
	$g = 0.5, h = 0.5$	0.0254	0.0258
	Likert-Moderate Skew	0.0400	0.0836
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.2230^a	
	Likert-Heavy Skew	0.0456	0.1344
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9570^a	
(30, 40, 50) / (1, 1, 16)	$g = 0.5, h = 0.0$	0.0460	0.0798
	$g = 1.0, h = 0.0$	0.0790	0.1212
	$g = 0.5, h = 0.5$	<u>0.0240</u>	0.0262
	Likert-Moderate Skew	0.0436	0.4360
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.2176^a	
	Likert-Heavy Skew	0.0558	0.5648
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9398^a	
(20, 40, 60) / (1, 1, 4)	$g = 0.5, h = 0.0$	0.0274	0.0334
	$g = 1.0, h = 0.0$	0.0442	<u>0.0538</u>
	$g = 0.5, h = 0.5$	<u>0.0208</u>	<u>0.0216</u>
	Likert-Moderate Skew	0.0206	0.0642
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.1818^a	
	Likert-Heavy Skew	0.0288	0.1124
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9658^a	

Table 8, continued.

Sample Sizes/Variances	Distribution	GLIMMIX	
		Baseline	Lognormal
(20, 40, 60) / (1, 1, 16)	$g = 0.5, h = 0.0$	0.0252	0.0478
	$g = 1.0, h = 0.0$	0.0542	0.0892
	$g = 0.5, h = 0.5$	<u>0.0124</u>	<u>0.0130</u>
	Likert-Moderate Skew	<u>0.0170</u>	0.3738
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.1762^a	
	Likert-Heavy Skew	0.0322	0.5150
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9410^a	

Note: ^aGLIMMIX modelled by multinomial response variable

Table 9. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method = ML) and various versions of GLIMMIX on symmetric data when group sample sizes are negatively paired with group variances

Sample Size/Variance	Distribution	GLIMMIX				
		A-D Normal Test				Lognormal
		Baseline	$\alpha=0.15$	$\alpha=0.20$	Akaike	
(30, 40, 50) / (1, 1, 4)	$g = 0.000, h = 0.000$	0.1024	0.1194	0.1230	0.1454	0.1044
	$g = 0.000, h = 0.225$	0.0922	0.0944	0.0950	0.0928	0.0964
	(0.975)N(0, 1) + (0.025)N(0, 25)	0.0992	0.1268	0.1280	0.1296	0.0976
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.0946	0.1140	0.1138	0.1148	0.0952
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.0908	0.0968	0.0968	0.0964	0.0894
	(0.975)N(0, 1) + (0.025)N(0, 100)	0.0714	0.1238	0.1242	0.1248	0.0650
	(0.950)N(0, 1) + (0.050)N(0, 100)	0.0736	0.1036	0.1038	0.1040	0.0668
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.0834	0.0804	0.0804	0.0804	0.0714
	Likert-Even	0.0970	0.2336	0.2336	0.0984	0.2028
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.1244^a				
	Likert-Symmetric	0.0984	0.0788	0.0788	0.0968	0.2070
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.0866^a				
	$g = 0.000, h = 0.000$	0.1450	0.1804	0.1866	0.2698	0.1852
	$g = 0.000, h = 0.225$	0.1402	0.1854	0.1856	0.1866	0.1480
(30, 40, 50) / (1, 1, 16)	(0.975)N(0, 1) + (0.025)N(0, 25)	0.1430	0.2384	0.2404	0.2538	0.1420
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.1420	0.2300	0.2302	0.2348	0.1432
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.1398	0.2102	0.2102	0.2110	0.1402
	(0.975)N(0, 1) + (0.025)N(0, 100)	0.1172	0.2398	0.2410	0.2426	0.1160
	(0.950)N(0, 1) + (0.050)N(0, 100)	0.1252	0.2178	0.2180	0.2180	0.1100
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.1378	0.1798	0.1798	0.1796	0.1086
	Likert-Even	0.1430	0.3624	0.3624	0.3624	0.3280
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.1866^a				
	Likert-Symmetric	0.1490	0.0462	0.0462	0.0462	0.1636
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.1140^a				

GENERALIZED LINEAR MODEL ANALYSES

Table 9, continued.

Sample Size/Variance	Distribution	GLIMMIX				
		Baseline	A-D Normal Test		Akaike	Lognormal
			$\alpha=0.15$	$\alpha=0.20$		
(20, 40, 60) / (1, 1, 4)	$g = 0.000, h = 0.000$	0.1110	0.1346	0.1372	0.1110	0.1158
	$g = 0.000, h = 0.225$	0.1060	0.0950	0.0948	0.0976	0.1082
	(0.975)N(0, 1) + (0.025)N(0, 25)	0.1620	0.1744	0.1738	0.1714	0.1578
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.1364	0.1616	0.1620	0.1626	0.1342
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.1386	0.1360	0.1362	0.1362	0.1426
	(0.975)N(0, 1) + (0.025)N(0, 100)	0.1376	0.1656	0.1656	0.1658	0.1276
	(0.950)N(0, 1) + (0.050)N(0, 100)	0.0992	0.1506	0.1506	0.1508	0.0860
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.1266	0.1140	0.1140	0.1140	0.1102
	Likert-Even	0.1450	0.2836	0.2836	0.1462	0.2380
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.1664^a				
	Likert-Symmetric	0.1472	0.0952	0.0952	0.1422	0.2546
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.1146^a				
	$g = 0.000, h = 0.000$	0.1720	0.2186	0.2258	0.1830	0.1996
	$g = 0.000, h = 0.225$	0.1686	0.1910	0.1910	0.1940	0.1650
(20, 40, 60) / (1, 1, 4)	(0.975)N(0, 1) + (0.025)N(0, 25)	0.2578	0.3156	0.3160	0.3276	0.2570
	(0.950)N(0, 1) + (0.050)N(0, 25)	0.2440	0.3128	0.3132	0.3174	0.2456
	(0.900)N(0, 1) + (0.100)N(0, 25)	0.2542	0.2852	0.2858	0.2856	0.2544
	(0.975)N(0, 1) + (0.025)N(0, 100)	0.2448	0.3198	0.3202	0.3206	0.2092
	(0.950)N(0, 1) + (0.050)N(0, 100)	0.2146	0.3014	0.3018	0.3020	0.1752
	(0.900)N(0, 1) + (0.100)N(0, 100)	0.2582	0.2550	0.2550	0.2552	0.1860
	Likert-Even	0.2328	0.4850	0.4850	0.4850	0.3728
	(0.2, 0.2, 0.2, 0.2, 0.2)	0.2532^a				
	Likert-Symmetric	0.2406	0.1056	0.1056	0.1062	0.2426
	(0.1, 0.2, 0.4, 0.2, 0.1)	0.1506^a				

Note: ^aGLIMMIX modelled by multinomial response variable

Table 10. Type I error rates for treatment group differences using the baseline test (GLIMMIX dist = normal method = ML) and GLIMMIX with lognormal response on skewed data when group sample sizes are negatively paired with group variances

Sample Sizes/Variances	Distribution	GLIMMIX	
		Baseline	Lognormal
(30, 40, 50) / (4, 1, 1)	$g = 0.5, h = 0.0$	0.1046	0.1132
	$g = 1.0, h = 0.0$	0.1102	0.1190
	$g = 0.5, h = 0.5$	0.0676	0.0692
	Likert-Moderate Skew	0.1030	0.1378
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.2962^a	
	Likert-Heavy Skew	0.1054	0.1822
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9240^a	

Table 10, continued.

Sample Sizes/Variates	Distribution	GLIMMIX	
		Baseline	Lognormal
(30, 40, 50) / (16, 1, 1)	$g = 0.5, h = 0.0$	0.1518	0.1894
	$g = 1.0, h = 0.0$	0.1906	0.2402
	$g = 0.5, h = 0.5$	0.1162	0.1154
	Likert-Moderate Skew	0.1456	
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.2944^a	0.5002
	Likert-Heavy Skew	0.1618	
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9152^a	0.5980
(30, 40, 50) / (16, 1, 1)	$g = 0.5, h = 0.0$	0.0460	0.0798
	$g = 1.0, h = 0.0$	0.0790	0.1212
	$g = 0.5, h = 0.5$	<u>0.0240</u>	0.0262
	Likert-Moderate Skew	0.0436	0.4360
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.2176^a	
	Likert-Heavy Skew	0.0558	0.5648
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.9398^a	
(20, 40, 60) / (4, 1, 1)	$g = 0.5, h = 0.0$	0.1126	0.1194
	$g = 1.0, h = 0.0$	0.1130	0.1218
	$g = 0.5, h = 0.5$	0.0714	0.0720
	Likert-Moderate Skew	0.1442	0.1722
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.3230^a	
	Likert-Heavy Skew	0.1512	0.2132
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.8850^a	
(20, 40, 60) / (16, 1, 1)	$g = 0.5, h = 0.0$	0.1796	0.2084
	$g = 1.0, h = 0.0$	0.2164	0.2510
	$g = 0.5, h = 0.5$	0.1286	0.1274
	Likert-Moderate Skew	0.2366	0.5066
	(0.50, 0.30, 0.15, 0.04, 0.01)	0.3242^a	
	Likert-Heavy Skew	0.2536	0.5870
	(0.70, 0.20, 0.06, 0.03, 0.01)	0.8844^a	

Note: ^aGLIMMIX modelled by multinomial response variable

Invited Article

Comparisons of Two Quantile Regression Smoothers

Rand R. Wilcox

University of Southern California
Los Angeles, CA

The small-sample properties of two non-parametric quantile regression estimators are compared. The first is based on constrained B-spline smoothing (COBS) and the other is based on a variation and slight extension of a running interval smoother. R functions for applying the methods were used in conjunction with default settings for the various optional arguments. Results indicate that the modified running interval smoother has practical value. Manipulation of the optional arguments might impact the relative merits of the two methods, but the extent to which this is the case remains unknown.

Keywords: Running interval smoother, COBS, Harrell-Davis estimator, LOWESS, Well Elderly 2 study, depressive symptoms, perceived control

---

Introduction

Consider the problem of estimating and plotting a regression line when the goal is to determine the conditional quantile of some random variable Y given X . Quantile regression methods have been studied extensively and plots of the regression line can provide a useful perspective regarding the association between two variables. One approach is to assume that the conditional q^{th} quantile of Y , given X , is given by

$$Y_q = \beta_0 + \beta_1 X, \quad (1)$$

where β_0 and β_1 are unknown parameters. For the special case where the goal is to estimate the median of Y , given X , least absolute regression can be used, which predates least squares regression by about a half century. A generalization, aimed

---

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at: rwilcox@usc.edu.

at dealing with any quantile, was derived by Koenker and Bassett (1978). While the assumption of a straight regression line appears to provide a good approximation of the true regression line in various situations, this is not always the case. One strategy for dealing with any possible curvature is to use some obvious parametric model. For example, add a quadratic term. But generally this can be unsatisfactory, which has led to the development of nonparametric regression lines, often called smoothers (e.g., Härdle, 1990; Efromovich, 1999; Eubank, 1999; Györfi, Kohler, Krzyżk, & Walk, 2002). For the particular case where the goal is to model the conditional quantile of Y , given X , one way of dealing with curvature in a reasonably flexible manner is to use constrained B-spline smoothing (COBS). The many computational details are summarized in Koenker and Ng (2005); see in particular section 4 of their paper. The Koenker–Ng method improves on a computational method studied by He and Ng (1999), and builds upon results in Koenker, Ng, and Portnoy (1994). Briefly, let $\rho_q(u) = u(q - I(u < 0))$, where the indicator function $I(u < 0) = 1$ if $u < 0$; otherwise $I(u < 0) = 0$. The goal is to estimate the q^{th} quantile of Y given X by finding a function $\omega(X)$ that minimizes

$$\sum_i \rho_q(Y_i - \omega(X_i)) \quad (2)$$

based on the random sample $(X_1, Y_1), \dots, (X_n, Y_n)$. The estimate is based on quadratic B-splines with the number of knots chosen via a Schwartz-type information criterion. Here, COBS is applied via the R package `cobs`.

The motivation for this study stems from the use of COBS when analyzing data from the Well Elderly 2 study (Jackson et al., 2009; Clark et al., 2012). A general goal was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study dealt with understanding the association between cortisol and various measures of stress and wellbeing. Before and six months following the intervention, participants were asked to provide, within 1 week, four saliva samples over the course of a single day, to be obtained on rising, 30 min after rising but before taking anything by mouth, before lunch, and before dinner. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicated measures of stress are associated with the cortisol awakening response, which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep. CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake for about an hour.

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

After intervention (with a sample size of 328), COBS indicated some seemingly unusually shaped regression lines. One of these had to do with the association between CAR and a measure of depressive symptoms using the Center for Epidemiologic Studies Depressive Scale (CESD). The CESD (Radloff, 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn, Hoberman, & Rosenbaum, 1988; Foley, Reed, Mutran, & DeVellis, 2002). Higher scores indicate a higher level of depressive symptoms. Figure 1 shows the estimated regression line for males when $q = 0.5$. (There were 157 males.) The estimated regression line for $q = 0.75$ had a shape very similar to the one shown in Figure 1.

Another portion of the study dealt with the association between CAR and a measure of perceived control. Perceived control was measured with the instrument in Eizenman, Nesselrode, Featherman, and Rowe (1997). The scores ranged between 16 and 32 and consisted of a sum of Likert scales. Now the 0.75 quantile regression line appears as shown in Figure 2. Again, there was concern about the shape of the regression line.

Figure 1. COBS regression line for predicting the 0.5 quantile of CESD, for males, based on the cortisol awakening response after intervention

Figure 2. COBS regression line for predicting the 0.75 quantile of perceived control based on the cortisol awakening response

One possibility is that the regression lines in [Figures 1 and 2](#) are a reasonable approximation of the true regression. But another possibility is that they reflect a type of curvature that poorly approximates the true regression line. Suppose that

$$Y = \beta_0 + \beta_1 X + \lambda(X)\varepsilon, \quad (3)$$

where $\lambda(X)$ is some function used to model heteroscedasticity and ε is a random variable having mean zero and variance σ^2 . Some preliminary simulation results suggested that if $\beta_0 = 0, \beta_1 = 1$, and both X and ε have standard normal distributions, reasonably straight regression lines are obtained using COBS. However, if ε has a skewed light-tailed distribution (a g -and- h distribution, details of which are described in a later section) and if, for example, $\lambda(X) = |X| + 1$, instances are encountered where a relatively high degree of curvature is encountered. An example is given in [Figure 3](#) with $n = 100$.

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

These results motivated consideration of an alternative quantile regression estimator. A few checks suggested that the problems just illustrated are reduced considerably, but there are no systematic simulation results providing some sense of how this alternative estimator compares to COBS. Consequently, the goal in this paper is to compare these estimators in terms of bias and mean squared error. Two additional criteria are used. The first is the maximum absolute error between the predicted and actual quantile being estimated. The other is aimed at characterizing how the estimators compare in terms of indicating a monotonic association when in fact one exists. This is done via Kendall's tau between the predicted and true quantiles.

Figure 3. COBS regression line for predicting the 0.5 quantile using generated data, $n = 100$

It is noted that COBS is being applied using the R package `cobs` in conjunction with default settings for the various arguments. The argument `lambda` alters how the regression line is estimated and might possibly improve the fit to data via visual inspection. But obviously this strategy is difficult to study via simulations. The

alternative estimator used here is applied with an R function (`qhds`), again using default settings for all of the arguments. The performance of the method is impacted by the choice for the span (the constant f in later sections). The simulations reported here provide information about the relative merits of the two estimators with the understanding that perhaps their relative merits might be altered based on a judgmental process that goes beyond the scope of this paper.

The following section provides the details of the alternative estimator. Later sections report simulation results comparing COBS to the alternative estimator and illustrate the difference between the two estimators for the data used in Figures 1-3.

Alternative Estimator

The alternative estimator consists of a blend of two smoothers: the running interval smoother (e.g., Wilcox, 2012) and the smoother derived by Cleveland (1979), typically known as LOWESS. The running interval has appeal because it is readily adapted to any robust estimator. In particular, it is easily applied when the goal is to estimate the conditional quantile of Y given X . However, often this smoother gives a somewhat jagged looking plot of the regression line. Primarily for aesthetic reasons, this issue is addressed by further smoothing the regression line via LOWESS.

The version of the running interval smoother used here is based in part on the quantile estimator derived by Harrell and Davis (1982). The Harrell–Davis estimate of the q^{th} quantile uses a weighted average of all the order statistics. Let $Z_1, \dots, Z_n$ be a random sample, let U be a random variable having a beta distribution with parameters $a = (n + 1)q$ and $b = (n + 1)(1 - q)$, and let

$$w_i = P\left(\frac{i-1}{n} \leq U \leq \frac{i}{n}\right).$$

The estimate of the q^{th} quantile is

$$\hat{\theta}_q = \sum w_i Z_{(i)},$$

where $Z_{(1)} \leq \dots \leq Z_{(n)}$ are the $Z_1, \dots, Z_n$ written in ascending order. Here the focus is on estimating the median and the 0.75 quantile. That is, $q = 0.5$ and 0.75 are used.

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

In terms of its standard error, Sfakianakis and Verginis (2006) show that the Harrell–Davis estimator competes well with alternative estimators that again use a weighted average of all the order statistics, but there are exceptions. Sfakianakis and Verginis derived alternative estimators that have advantages over the Harrell–Davis in some situations, but when sampling from heavy-tailed distributions, the standard error of their estimators can be substantially larger than the standard error of $\hat{\theta}_q$. Comparisons with other quantile estimators are reported by Parrish (1990), Sheather and Marron (1990), as well as Dielman, Lowry, and Pfaffenberger (1994). The only certainty is that no single estimator dominates in terms of efficiency. For example, the Harrell–Davis estimator has a smaller standard error than the usual sample median when sampling from a normal distribution or a distribution that has relatively light tails, but for sufficiently heavy-tailed distributions, the reverse is true (Wilcox, 2012, p. 87).

The running interval smoother is applied as follows: Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be a random sample from some unknown bivariate distribution, and let f be some constant to be determined. Then the point x is said to be close to X_i if

$$|X_i - x| \leq f \times \text{MADN} ,$$

where MADN is $\text{MAD}/0.6745$, MAD is the median of $|X_1 - M|, \dots, |X_n - M|$, and M is the usual sample median based on $X_1, \dots, X_n$. For normal distributions, MADN estimates the standard deviation, in which case x is close to X_i if x is within f standard deviations of X_i . Let

$$N(X_i) = \{j : |X_j - X_i| \leq f \times \text{MADN}\} .$$

That is, $N(X)$ indexes the set of all X_j values that are close to X_i . Let $\hat{\theta}_i$ be the Harrell–Davis estimate based on the Y_j values such that $j \in N(X_i)$. To get a graphical representation of the regression line, compute $\hat{\theta}_i$, the estimated value of Y given that $X = X_i$, $i = 1, \dots, n$, and then plot the points $(X_i, \hat{\theta}_1), \dots, (X_n, \hat{\theta}_n)$. Typically $f = 0.8$ or 1 gives good results, but of course exceptions are encountered. Here, $f = 0.8$ is assumed unless stated otherwise.

As previously indicated, the plot produced by the running interval smoother can be a bit ragged. Consequently, the initial smooth was smoothed again by proceeding as follows: Given X_j , let $\delta_i = |X_i - X_j|$, $i = 1, \dots, n$.

RAND R. WILCOX

Sort the δ_i values and retain the ξ_n pairs of points that have the smallest δ_i values, where ξ is a number between 0 and 1 and plays the role of a span. Here, $\xi = 0.75$ is used. Let δ_m be the largest δ_i value among the retained points. Let

$$Q_i = \frac{|X_j - X_i|}{\delta_m}$$

and, if $0 \leq Q_i < 1$, set

$$w_i = (1 - Q_i^3)^3 .$$

Otherwise, set $w_i = 0$. Next, use weighted least squares to predict $\hat{\theta}_j$ corresponding to X using the w_i values as weights. That is, determine the values b_1 and b_0 that minimize

$$\sum w_i (\hat{\theta}_i - b_0 - b_1 X_i)^2$$

and estimate $\hat{\theta}_j$ with $\tilde{\theta}_j = b_0 + b_1 X_j$. The final plot of the quantile regression is taken to be the line connecting the points $(X_j, \tilde{\theta}_j)^2$ ($j = 1, \dots, n$). This will be called method R henceforth.

Simulation

Simulations were used to compare the small-sample properties of COBS and the modified running interval smoother based on $K = 4000$ replications and sample size $n = 50$. The data were generated according to the model

$$Y = X + \lambda(X)\varepsilon , \tag{4}$$

where X is taken to have a standard normal distribution and ε has one of four distributions: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, the distribution for the error term was taken to be one of four g -and- h distributions (Hoaglin, 1985) that contain the

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

standard normal distribution as a special case. Let Z be a random variable that has a standard normal distribution, and let

$$W = \frac{\exp(gZ) - 1}{g} \exp\left(hZ^2/2\right)$$

unless $g = 0$, in which case

$$W = Z \exp\left(h \frac{Z^2}{2}\right).$$

Then W has a g -and- h distribution, where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0.0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Hoaglin (1985) summarizes additional properties of the g -and- h distributions.

Table 1. Some properties of the g -and- h distribution

g	h	k1	k2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Three choices for λ were considered: $\lambda \equiv 1$, $\lambda = |X| + 1$, and $\lambda = 1/(|X| + 1)$. These three choices are henceforth called VP 1, 2, and 3, respectively.

Note that based on how the data are generated, as indicated by (5), ideally a smoother should indicate a monotonic increasing association between $X_1, \dots, X_n$ and $\tilde{\theta}_1, \dots, \tilde{\theta}_n$, where $\tilde{\theta}_i$ is the estimate of the q^{th} quantile of Y , given that $X = X_i$, based on either COBS or method R. The degree to which this goal was accomplished was measured with Kendall's tau.

Details about the four criteria used to compare COBS and method R are as follows: The first criterion was mean squared error, which was estimated with

$$\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n (\theta_{ik} - \tilde{\theta}_{ik})^2, \quad (5)$$

where now, for the k^{th} replication, θ_{ik} is the true conditional q^{th} quantile of Y given $X = X_i$. Bias was estimated with

$$\frac{1}{nK} \sum_{k=1}^K \sum_{i=1}^n \theta_{ik} - \tilde{\theta}_{ik}. \quad (6)$$

The third criterion was the mean maximum absolute error:

$$\frac{1}{K} \sum_{k=1}^K \max \left\{ |\theta_{1k} - \tilde{\theta}_{1k}|, \dots, |\theta_{nk} - \tilde{\theta}_{nk}| \right\}. \quad (7)$$

The fourth criterion was

$$\frac{1}{K} \sum \tau_k, \quad (8)$$

where, for the k^{th} replication, τ_k is Kendall's tau between $X_1, \dots, X_n$ and $\tilde{\theta}_1, \dots, \tilde{\theta}_n$.

It is noted that the θ_{ik} values are readily determined because the transformation used to generate observations from a g -and- h distribution is monotonic and quantiles are location and scale equivariant.

Simulation results are reported in [Tables 2](#) and [3](#), where RMSE is the mean squared error of COBS divided by the mean squared error of method R, and RMAX is the maximum absolute value of COBS divided by the maximum absolute value of the error based on method R. As can be seen, generally method R competes well with COBS in terms of RMSE and RMAX, but neither method dominates. For $q = 0.5$, R is uniformly better in terms of RMSE, but for $q = 0.75$ and VP 3, COBS performs better than R. As for RMAX, R performs best for VP 1 and 2, while for VP 3 the reverse is true. Bias for both methods is typically low, with COBS seeming to have an advantage over method R. The main result is that in terms of τ , method R dominates. That is, the simulations indicate that method R is better at avoiding an indication of curvature that does not reflect the true regression line, as was the case in [Figure 3](#).

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

Table 2. Simulation results for $q = 0.5$

g	h	VP	RMSE	RMAX	BIAS		r	
					COBS	R	COBS	R
0.0	0.0	1	1.284	1.293	0.002	0.002	0.957	0.997
0.0	0.0	2	1.405	2.050	-0.002	-0.002	0.773	0.927
0.0	0.0	3	1.281	0.726	-0.001	-0.001	0.989	1.000
0.0	0.2	1	1.160	1.333	-0.002	-0.002	0.955	0.994
0.0	0.2	2	1.395	2.076	0.005	0.009	0.794	0.917
0.0	0.2	3	1.104	0.753	-0.003	-0.002	0.991	1.000
0.2	0.0	1	1.247	1.292	0.015	0.031	0.954	0.996
0.2	0.0	2	1.400	2.048	0.023	0.035	0.786	0.930
0.2	0.0	3	1.220	0.732	0.004	0.022	0.989	1.000
0.2	0.2	1	1.178	1.384	0.014	0.027	0.956	0.993
0.2	0.2	2	1.455	2.155	0.034	0.042	0.794	0.914
0.2	0.2	3	1.040	0.765	0.005	0.023	0.990	1.000

Table 3. Simulation results for $q = 0.75$

g	h	VP	RMSE	RMAX	BIAS		r	
					COBS	R	COBS	R
0.0	0.0	1	1.046	1.375	-0.017	0.077	0.938	0.994
0.0	0.0	2	1.328	2.020	-0.052	0.027	0.709	0.862
0.0	0.0	3	0.644	0.807	-0.014	0.105	0.973	0.998
0.0	0.2	1	0.847	1.459	0.010	0.137	0.911	0.978
0.0	0.2	2	1.124	2.140	0.022	0.136	0.666	0.794
0.0	0.2	3	0.544	0.858	0.001	0.145	0.969	0.995
0.2	0.0	1	0.964	1.423	0.000	0.110	0.907	0.985
0.2	0.0	2	1.284	2.057	-0.026	0.074	0.655	0.803
0.2	0.0	3	0.642	0.860	-0.006	0.126	0.962	0.997
0.2	0.2	1	0.849	1.505	0.042	0.181	0.880	0.953
0.2	0.2	2	1.195	2.214	0.084	0.202	0.614	0.740
0.2	0.2	3	0.552	0.945	0.013	0.167	0.955	0.993

Illustrations

The data in Figures 1-3 are used to illustrate method R. The left panel of Figure 4 shows the 0.5 quantile regression line for CAR and CESD. Notice that, for CAR positive (cortisol decreases after awakening), the plot suggests a positive association with depressive symptoms, which is consistent with Figure 1. But, for CAR negative, method R suggests that there is little or no association with CESD and clearly provides a different sense regarding the nature of the association. A criticism might be that, if method R were to use a smaller choice for the span, perhaps an association similar to Figure 1 would be revealed. But even with a span of $f=0.5$, the plot of the regression line is very similar to the one shown in Figure 4.

The right panel of Figure 4 shows the 0.75 quantile regression line for predicting perceived control based on CAR, which differs in an obvious way from the regression line based on COBS shown in Figure 2. Figure 4 indicates that there is little or no indication of an association with CAR when CAR is negative, but for CAR positive, a negative association is indicated. The only point is that the choice between COBS and method R can make a substantial difference.

Figure 5 shows the 0.5 quantile regression line based on the data used in Figure 3. In contrast to COBS, method R provides a very good approximation of the true regression line. Again, this only illustrates the extent to which the two methods can give strikingly different results. As is evident, in this particularly case, method R provides a much more accurate indication of the true regression line.

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

Figure 4. The quantile regression lines using method R and the data in [Figures 1 and 2](#)

Figure 5. The quantile regression line using method R and the data in [Figure 3](#)

Conclusion

For the situations considered in the simulations, method R does not dominate COBS based on the four criteria used here. COBS seems to have an advantage in terms of minimizing bias. But otherwise, method R competes well with COBS, particularly in terms of Kendall's tau, which suggests that typically method R is better able to avoid an indication of spurious curvature. Moreover, the illustrations demonstrate that the choice between the two methods can make a substantial difference even with a sample size of $n = 328$. So in summary, method R would seem to deserve serious consideration.

Another possible appeal of method R is that it is readily extended to the situation where there is more than one independent variable. That is, a generalization of the running interval smooth already exists (e.g., Wilcox, 2012). Moreover, additional smoothing can be accomplished, if desired, using the smoother derived by Cleveland and Devlin (1988), which generalizes the technique derived by Cleveland (1979). Evidently, a generalization of COBS to more than one independent variable has not been derived.

Finally, an R function for applying method R, called `ghdsm`, is available in the R package WRS, a version of which is provided as a supplemental item to this article. Updated versions may be found at dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm under the "Software" tab.

References

- Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology*, 80(3), 265-278. doi: 10.1016/j.biopsycho.2008.10.004
- Clark, F., Jackson, J., Carlson, M., Chou, C., Cherry, B. J., Jordan-Marsh, M.,...Azen, S. P. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, 66(9), 782-790. doi: 10.1136/jech.2009.099754
- Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829-836. doi: 10.1080/01621459.1979.10481038

COMPARISONS OF TWO QUANTILE REGRESSION SMOOTHERS

- Cleveland, W.S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596-610. doi: [10.1080/01621459.1988.10478639](https://doi.org/10.1080/01621459.1988.10478639)
- Clow, A., Thorn, L., Evans, P., & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress: The International Journal on the Biology of Stress*, 7(1), 29-37. doi: [10.1080/10253890410001667205](https://doi.org/10.1080/10253890410001667205)
- Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics – Simulation and Computation*, 23(2), 355-371. doi: [10.1080/03610919408813175](https://doi.org/10.1080/03610919408813175)
- Efromovich, S. (1999). *Nonparametric curve estimation: Methods, theory and applications*. New York, NY: Springer-Verlag.
- Eizenman, D. R., Nesselroade, J. R., Featherman, D. L., & Rowe, J. W. (1997). Intraindividual variability in perceived control in an older sample: The MacArthur successful aging studies. *Psychology and Aging*, 12(3), 489-502. doi: [10.1037/0882-7974.12.3.489](https://doi.org/10.1037/0882-7974.12.3.489)
- Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York, NY: Marcel Dekker.
- Foley K., Reed P., Mutran, E., & DeVellis, R. F. (2002). Measurement adequacy of the CES-D among a sample of older African-Americans. *Psychiatric Research*, 109(1), 61-69. doi: [10.1016/S0165-1781\(01\)00360-2](https://doi.org/10.1016/S0165-1781(01)00360-2)
- Györfi, L., Kohler, M., Krzyzk, A., & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York, NY: Springer Verlag.
- Härdle, W. (1990). *Applied nonparametric regression*. Econometric Society Monographs (No. 19). Cambridge, UK: Cambridge University Press.
- Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69(3), 635-640. doi: [10.1093/biomet/69.3.635](https://doi.org/10.1093/biomet/69.3.635)
- He, X., & Ng, P. (1999). COBS: Qualitative constrained smoothing via linear programming. *Computational Statistics*, 14(3), 315-337. doi: [10.1007/s001800050019](https://doi.org/10.1007/s001800050019)
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring Data Tables Trends and Shapes* (pp. 461-515). New York, NY: Wiley.
- Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S.,...Clark, F. (2009). Confronting challenges in intervention research with

- ethnically diverse older adults: The USC Well Elderly II trial. *Clinical Trials*, 6(1), 90-101. doi: 10.1177/1740774508101191
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33-50. doi: 10.2307/1913643
- Koenker, R., & Ng, P. (2005). Inequality constrained quantile regression. *Sankhyā: The Indian Journal of Statistics*, 67(2), 418-440. Available from <http://www.jstor.org/stable/25053440>
- Koenker, R., Ng, P., & Portnoy, S. (1994). Quantile smoothing splines. *Biometrika*, 81(4), 673-680. doi: 10.1093/biomet/81.4.673
- Lewinsohn, P. M., Hoberman, H. M., & Rosenbaum M. (1988). A prospective study of risk factors for unipolar depression. *Journal of Abnormal Psychology*, 97(3), 251-264. doi: 10.1037/0021-843X.97.3.251
- Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, 46(1), 247-257. doi: 10.2307/2531649
- Radloff, L. (1977). The CES-D scale: A self report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3), 385-401. doi: 10.1177/014662167700100306
- Sfakianakis, M. E., & Verginis, D. G. (2006). A new family of nonparametric quantile estimators. *Communications in Statistics – Simulation and Computation*, 37(2), 337-345. doi: 10.1080/03610910701790491
- Sheather, S. J., & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410), 410-416. doi: 10.1080/01621459.1990.10476214
- Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). San Diego, CA: Academic Press.

New Nonparametric Rank Tests for Interactions in Factorial Designs with Repeated Measures

Jos Feys

University of Leuven
Leuven, Belgium

New rank tests for interactions in factorial designs are presented and applied to some common factorial designs with repeated measures. The resulting p -values of these tests are compared, along with those obtained by parametric and randomization tests.

Keywords: Nonparametric interaction tests, aligned rank test, new rank-based methods, Friedman ranks, randomization, gain scores

---

Introduction

Techniques have been proposed for the nonparametric analysis of interactions in factorial designs. They rank the observations and then perform parametric tests on ranks. The aligned rank tests belong to one class of these. Aligning implies that some estimate of a location (e.g., for the effect on a certain level of a given factor), such as the mean or median of the observation, is subtracted from each observation. These data, thus aligned according to the desired main and/or interaction effects, are then ranked and parametric tests are performed on these aligned ranks. A second class of such tests, named the new rank based methods, first rank the (not aligned) observations, and then the relative treatment effects are defined in reference to the distribution of the variables measured and estimated through elaborate calculations on these ranks.

The alignment methodology was introduced by Hodges and Lehmann (1962) and extended to two-way layouts by Sen (1968). McSweeney (1967) developed a test (M test) for interaction using the aligned ranks in the two-way layout. The aligned rank tests were publicized by: Hettmansperger (1984), Puri and Sen (1985), Sawilowsky (1990), and Higgins and Tashtoush (1994). More recently, Beasley

---

Jos Feys is an Honorary Senior Research Fellow in the department of Kinesiology. Email at: jos.feys@faber.kuleuven.be.

and Zumbo (2003; 2009) added the aligned Friedman rank test for interactions in split-plot or repeated measures designs. Reviews of the aligned rank tests have been provided by Sawilowsky (1990), Higgins and Tashtoush (1994), Toothaker and Newman (1994), Kelley and Sawilowsky (1997), Richter and Payton (1999), Peterson (2002), and Rodriguez, Álvarez, and Ramirez (2009). Salazar-Álvarez, Tercero-Gómez, Temblador-Pérez, and Conover (2014) recently reviewed nonparametric test for interactions. They overlooked the – in my opinion – important contributions by Beasley and Zumbo (2003; 2009) on designs with repeated measures. The general conclusion of these reviews was that the aligned rank tests are valid nonparametric alternatives for the parametric tests for the interaction, especially when sample sizes are small (Sawilowsky, 1990) or the departure from normality of the distribution of the observations is extreme (e.g., heavy tailed; Kelley & Sawilowsky, 1997).

Pioneers on the new rank-based methods were Akritas (1990), Akritas and Arnold (1994), Akritas, Arnold, and Brunner (1997), and Akritas and Brunner (1997). Brunner and Puri (2001) reviewed these methods.

New Rank Tests

The Aligned Rank Transform Formulae

The rank transform procedure, as proposed by Conover and Iman (1976), in essence replaces original observations with their ranks and then computes parametric tests on these ranks. Higgins and Tashtoush (1994) showed that this method is flawed when applied to tests for interaction in factorial designs. The underlying reason is that, when nonlinear transformations (such as the rank transform) are made on a set of data, interaction structures that exist may or may not exist in the transformed data, and vice versa. Therefore, the rank transform procedure cannot be applied to test interactions. They advocated the use of the alignment of the data before ranking, the aligned rank transform, thus combining the notion of alignment of data and the rank transform. This procedure removes the effect of nuisance – as they called it – parameters when testing for effects of parameters of interest.

Two-way between designs

For two-way designs ($A \times B$, 2 between factors), the mathematical linear model is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} , \quad (1)$$

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

where $i = 1, \dots, r$, $j = 1, \dots, c$, $k = 1, \dots, n$, and the ϵ_{ijk} s are independent and identically distributed (i.i.d.) random variables with mean 0 and common standard deviation σ is (Higgins & Tashtoush, 1994, p. 203). The α_i s and β_j s represent the row (A) and column (B) effect, respectively, and $(\alpha\beta)_{ij}$ is the interaction. The adjustment factors proposed are based upon the usual estimates of the parameters. These estimates, with their respective means, are

$$\hat{\mu} = \bar{Y}, \quad \hat{\alpha}_i = \bar{Y}_i - \bar{Y}, \quad \hat{\beta}_j = \bar{Y}_j - \bar{Y}, \quad \hat{\alpha\beta}_{ij} = \bar{Y}_{ij} - \bar{Y}_i - \bar{Y}_j + \bar{Y}.$$

The aligned data for testing $A \times B$ interactions have the form

$$AB_{ijk} = Y_{ijk} - \bar{Y}_i - \bar{Y}_j + \bar{Y}. \quad (2)$$

The row (A) and column (B) effects (means) are subtracted from the individual observations and the overall mean is added (to compensate for the 2 subtractions). To apply the aligned rank transform to test for interactions, the AB_{ijk} s are ranked, and the ranked data are analyzed with a full model parametric procedure which includes all effects (i.e. effect A, effect B, and effect $A \times B$), for example an ANOVA. The authors compared the use of the sample means, as estimates of the location of the effects, to the use the medians and trimmed means. They advocated the use of the means because it is easy to calculate and is equally powerful as its alternatives. Peterson (2002) compared six alternatives for the estimation of location and concluded that the samples means and medians are the best estimators.

Repeated measures

For split-plot or repeated measures designs, the aligned data for testing for interactions is given by Higgins and Tashtoush (1994, p. 208):

$$GT_{ijk} = Y_{ijk} - \bar{Y}_{i.k} - \bar{Y}_{.j} + \bar{Y}, \quad (3)$$

where $i = 1, \dots, r$ (between Groups), $j = 1, \dots, c$ (within Time), $k = 1, \dots, n$ (observations/subjects/blocks), and GT_{ijk} represents the Group $\times$ Time interaction. The mean $\bar{Y}_{i.k}$ of observations of the k^{th} row (observation nested within the i^{th} Group) and the mean $\bar{Y}_{.j}$ of observations of the j^{th} column (Time) are subtracted from the individual observation Y_{ijk} , and the overall mean is added to the difference.

Three-way designs

For three-way interactions in experiments involving three factors, the alignment is given schematically as

$$\begin{aligned}
 ABC_{ijkl} = & Y_{ijkl} \\
 & - (\text{sum of 2-way means involving } i, j, k) \\
 & + (\text{sum of 1-way means involving } i, j, k) \\
 & - \text{overall mean}
 \end{aligned} \tag{4}$$

(Higgins & Tashtoush, 1994, p. 209). These formulae can be applied with a little programming skill in a data manipulation step of any kind of statistical package. Once this is done, ranking and testing for interactions should be a routine job for everyone. SAS statistical software (SAS Institute Inc., 2008) is used here for implementing these formulae in all subsequent examples. Leys and Schumann (2010) used (2) for the $A \times B$ interaction above on ordered Likert-type scale data.

Randomization or Permutation

Randomization or resampling tests have become popular since they were included in some procedures or functions within common statistical software. Cassell (2002) developed randomization test wrapper macros for the SAS statistical software procedures. With these macros wrapped around common parametric procedures, these can be replicated for a large number of random data permutations. Then the number of times the obtained p -values are equal to or smaller than the parametric test p -value are counted. The result of this count divided by the number of random permutations (usually 10000) is the randomization test p -value. In the examples below, I used this randomization test wrapper for comparison with the parametric and nonparametric procedures.

New Rank-Based Methods, Brunner SAS and R Macros, Gao and Alvo

Brunner, Domhof, and Langer (2002) developed macros with SAS and R (R Core Team, 2012) for the applications of the new rank-based methods. Noguchi, Gel, Brunner, and Konietzschke (2012) published the nparLD R package which provides researchers an easy and user-friendly access to these methods. Along with a macro for calculating confidence intervals, the macros offered range from a within

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

(repeated measures or longitudinal) one-way design and a between one-way design, up to three-way designs with one or two within factors. The original data are always first ranked. All macros accept data only in long (or multiple record) format. Therefore, if the data are in short (or multiple variable) format, with one record per observation (per row) and the repeated measures in the columns, they are stretched out to separate records (rows) for each repeated measure and only one response (dependent variable), and a variable is added to identify the repeated measure. In each example below, an appropriate macro was used.

Shah and Madden (2004), who illustrated the usefulness of several of the SAS macros in plant disease epidemiology, also presented the theory in a succinct but clear way. In this methodology, based upon the seminal paper by Akritas and Arnold (1994), the hypotheses are not formulated in terms of expectations of treatment effects (e.g., difference between means), but rather in reference to the distribution of variables measured in the experiment. Marginal or treatment (the authors use both terms interchangeably) effects are quantified by the appropriate estimates on mean ranks after extensive calculations. Noguchi et al. (2012) noted “the rank-based methodology is not restricted to data on a continuous scale and enables to analyze ordered categorical, dichotomous, and heavily skewed data,” (p. 2). They added that the methods are robust to outliers and are appropriate for small sample sizes. Akritas et al. (1997) specified that these methods are suitable for unbalanced designs. Kaptein, Nass, and Markopoulos (2010) demonstrated the power of this approach for the analysis of Likert-type rating scales.

In an electronic supplement, Shah and Madden (2004) explained how to apply the new rank-based methods using SAS for mixed (i.e., fixed and random effects) models. The data are not aligned; the MIXED procedure is applied directly on ranked data. It allows different covariance structures for all factor level combinations by specifying the type = UN (unstructured variance-covariance) option. The use of the so-called minimum variance quadratic unbiased estimation method is recommended instead of the default restricted maximum likelihood. We used this procedure in Example 1.

Gao and Alvo (2005a; 2005b) added a new rank statistic to test for interactions in two-way layouts by comparing the sum of row ranks with the sum of column ranks. It is unclear how to apply this statistic to two-way layouts with repeated measures.

Aligned Friedman Ranks, Beasley and Zumbo SAS/IML

Beasley and Zumbo (2003; 2009) investigated the usefulness of aligned rank tests for interactions in split-plot or repeated measures designs. The alignment procedure they advocated for such designs is the Higgins and Tashtoush (1994; hereafter called H&T) method whereby the alignment takes into account that the observations (subjects/blocks) are nested within the between factor Group. The aligned data for testing the interaction are obtained by applying the above mentioned (3) for split-plot or repeated measures designs. The authors methodically compared the regular rank test, across observations, on the aligned data with the aligned Koch ranks based on ranking the K^2 pairwise differences among the K levels of the repeated measures, regardless of Group membership, and aligned Friedman ranks, based on ranking of the data from 1 to K across the levels of the repeated measures factor within each observation. In their 2009 article, they included the SAS/IML (Interactive Matrix Language) syntax code to perform the aligned regular rank, the aligned Friedman rank, and the aligned Koch rank tests. The data are first aligned and then ranked according to each of these methods. After these two steps, parametric procedures with all effects included in the model (i.e., a full-factorial model repeated measures ANOVA) are applied to the three versions of aligned ranks. In most examples below, all with repeated measures, this IML script was applied to obtain the desired aligned ranks tests along with calculations based on the H&T formulae as checks.

ARTool

Wobbrock, Findlater, and Higgins (2011) proposed the ARTool, a tool for calculating the aligned rank transform. The ARTool generalizes aligned rank transform for nonparametric factorial data analysis to N factors and can therefore be used for higher-order interactions. The alignment of the data (in long format: with the observation identifier in the first column and the response in the last column, and all intervening factors in-between) is made in five steps. First (step 1), the residuals (observations – cell mean) are computed, and then (step 2) the estimated effects for all main and interaction effects are computed. For example, for a two-way design, the estimated effect for an $A \times B$ interaction response is achieved by

$$Y_{ij} = \overline{A_i B_j} - \bar{A}_i - \bar{B}_j + \mu . \quad (5)$$

In a third step, the aligned response is computed:

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

$$Y' = \text{residual} + \text{estimated effect (result from step 1 + step 2)} .$$

These aligned data (Y') are ranked to create Y'' , the aligned ranks. Finally, a full-factorial ANOVA is performed on Y'' . Wobbrock et al. (2011) noted that “alignment works best for completely randomized designs; it also works for other designs, but effects may not be entirely stripped out” (p. 146). This application was used in all examples below.

Examples

Example 1: A Pretest-Posttest Design

The first example was taken from Bonate (2000, p. 106). The data in the table resulted from a design with two Groups (control and treatment; $n = 10$ and $n = 9$, respectively) and two repeated (pre- and post-) measures, denoted as Time factor. In the treatment Group, there was an outlier on the post-measures: a value of 19 between values quite larger than 60 in the whole table. With Dixon's test for a single outlier, this very low value was flagged with a test probability (one sided) of $p < 0.001$.

According to Bonate (2000), pretest-posttest data can be analyzed in several ways: ANOVA on final scores alone, on difference scores, on percentages change scores, by means of an analysis of covariance (ANCOVA) with the pretest as covariate for the predicting Group factor and the posttest as outcome variable, blocking by initial scores (stratification), and as repeated measures. In this example, focus on difference or gain scores, repeated measures, and ANCOVA. The resulting p -values for the interaction with the different ways of analysis are reported in Table 1.

Gain Scores

Gain scores are obtained by computing post – pre difference scores. Differences in gain scores (i.e., in difference scores) between Groups, if any, should reveal the interaction Group $\times$ Time. If there is more gain in one Group than in the other, this would correspond to the interaction.

Parametric tests. The distribution of the gain scores over the 2 Groups was not normal, according to the Shapiro-Wilk test, $p < 0.0122$. The means and standard deviations (between brackets) of the gain scores, in the control Group and in the

treatment Group respectively, were: -1.60 (7.31) vs. 15.56 (29.77). So, the difference in gain scores was: 17.16 (21.10) in favor of the treatment Group. This seemed to indicate that there was an interaction between the Time (within) and the Group (between) factor. The variances of the gain scores were not equal between Groups, $p = 0.0003$. The t-test on the difference in gain scores between Groups (i.e., Group $\times$ Time interaction), for equal variances (pooled), was not significant, $p = 0.0947$; for unequal variances, this test (Satterthwaite corrected df's) was also not significant, $p = 0.1270$.

Table 1. Resulting p -values with different way of analysis for the Group $\times$ Time interaction in Example 1.

Gain scores	p -values	Rep. measures	p -values	(R)ANCOVA	p -values
Parametric/randomization		Parametric/randomization		Parametric/randomization	
t-test		F-test		ANCOVA posttest,	
pooled	0.0947	short format	0.0947	pretest as covariate	
Satterthwaite	0.127	long format	0.0947	F-test	0.0576
Wrapper t-test		Wrapper F-test	0.0863	Wrapper F-test	0.0474
pooled	0.0862				
Satterthwaite	0.1202				
Permutation option,					
NPARIWAY	0.0892				
Nonparametric		Nonparametric		Nonparametric	
on ranked gain scores		on ranked pre-post measures		on residual ranks	
t-test, pooled	0.0015	F-test	0.0615	Quade's	
				RANCOVA	0.0048
Wrapper t-test	0.0022	new rank-based methods		Mant.-Heanszel	0.0088
Permutation option,		(Shah & Madden)		Wilcox. exact	0.0057
NPARIWAY	0.0025	F-appr. large N.	0.0484	Permutation options,	
		F-appr. small N	0.067	NPARIWAY	0.0061
on gain scores, ranked		on pre-post measures,		ranked posttest,	
in procedure		ranked in procedure		ranked pretest as covariate	
Mant.-Heanszel	0.0014	Brunner macro F1_LD_F1		RANCOVA	
Wilcox. exact	0.0027	F-appr. large N	0.0484	F-test	0.0031
Brunner OWL		F-appr. small N	0.067		
Exact test	0.0027	aligned ranks			
F-approximation	0.0017	regular	0.0014		
		Friedman	0.0011		
		Koch	0.0015		
		ARTool	< 0.0001		

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

The Cassell randomization wrapper around this t-test procedure, showed almost the same p -values as these two values for the equal or unequal variances, respectively: $p = 0.0862$ and $p = 0.1202$. The permutation test option in the SAS nonparametric one-way procedure (NPAR1WAY) on the gain score also revealed a non-significant p value: 0.0892.

Nonparametric tests. The gain scores were ranked over the 2 Groups, thus ignoring the Group factor, and several parametric tests were applied on these ranks. The Shapiro-Wilk test for normality of distribution on the ranks was not significant, $p = 0.5236$. The mean ranks (SD's) were, for the control Group: 6.50 (3.00) and, for the treatment Group: 13.89 (5.34). The variances between Groups were equal, $p = .1051$. The (pooled) t-test for the difference of 7.39 (4.27) between Groups (i.e., Group $\times$ Time interaction) was significant, $p = .0015$. The randomization wrapper around this t-test upon the ranked gain scores resulted in a quasi-equally significant value, $p = .0022$. The permutation test option in the SAS nonparametric one-way procedure also yielded a significant p -value for the difference in ranked gain scores, $p = .0025$.

Typical nonparametric tests can rank the scores within the procedure itself. One of these is the Mantel-Heanszel statistic on the differences between the two Groups in gain scores, ranked within the SAS FREQ procedure with the scores = ranks option. This was significant, $p = 0.0014$. The exact Wilcoxon two-sample test on gain scores (nonparametric one-way procedure) was also significant, $p = 0.0027$. The Brunner macro One-Way Layout (OWL) for the Group factor on the gain scores, ranked within this macro, resulted in similar p -values; the exact $p = 0.0027$, and the value for the F-approximation was $p = .0017$.

Repeated Measures

Parametric tests. The means and standard deviations (between brackets) of the pretest measures for the control Group and treatment Group, respectively, were about equal: 75.00 (4.50) vs. 78.78 (5.89). For the posttest measures, these values were quite different: 73.40 (7.09) vs. 94.33 (28.74), respectively. Again, as for the differences in gain scores, this seemed to indicate an interaction. For a Time (pre-post) $\times$ Groups design, a repeated measure ANOVA F-test for the interaction corresponds to t-test on difference or gain scores, as pointed out by Dimitrov and Rumrill (2003). The p -value for this F-test on the data in short format was the exactly the same as the p -value for the pooled (i.e. df's for equal variances) t-test, namely $p = 0.0947$. Of course, this p value for the F-test was also the same when

calculated on the data in long format. The randomization test with the Cassell wrapper around this repeated measures ANOVA procedure was also not significant for this interaction, F-test, $p = 0.0863$, which was about the same value as with the randomization wrapper pooled t-test on gain scores in Table 1.

Nonparametric tests. The pretest-posttest measures were ranked on the data in long format, thus over Groups and Times. The mean ranks and their standard deviations (between brackets) of the pretest measures for the control Group and treatment Group, respectively, were: 14.85 (7.93) vs. 20.11 (7.19). For the posttest measures, respectively: 13.45 (9.48) vs. 30.78 (11.40). The Shapiro-Wilk test for normality of distributions was not significant, $p = 0.1223$. The F-test for the Group $\times$ Time interaction on these ranks was not significant, $p = 0.0615$. Applying the parametric SAS procedure MIXED on these ranks to obtain the estimated treatment effect of the new rank-based methods, as described in the electronic supplement to Shah and Madden (2004), resulted in a significant p value for the F-approximation for large samples, $p = 0.0484$, but not for small samples, $p = 0.0670$.

Here too, some nonparametric procedures include ranking. The Brunner macro suited for this design (F1_LD_F1) was applied to the repeated measures data in long format, which yielded exactly the same p -values as obtained with the Shah and Madden (2004) parametric SAS procedure for large and small samples, respectively: 0.0485 and 0.0670.

The Beasley and Zumbo (2009) SAS/IML syntax code includes the calculations based on the H&T formula (3) aimed at the alignment for the interaction with split plot or repeated measures in short format. Applying this script (paralleled with the author's own calculations) yielded a significant p value for the regular ranking across observations of the aligned data: 0.0014. For the Friedman ranking (within observations), as well as for the Koch rankings, the resulting p -values were also quite significant, respectively: 0.0011 and 0.0015.

The ARTool can only be applied on data in long format. The ARTool procedure was applied to the data in this format and a significant p -value for the interaction was obtained, $p < 0.0001$. The Pearson correlation between the data aligned with the H&T formula and the ARTool method was: $r(N = 38) = 0.7386$, which indicated that the alignments were not the same.

Analysis of Covariance (ANCOVA)

The interaction Group $\times$ Time can also be tested with an ANCOVA test by specifying the pretest measure to be the covariate and the Group variable to be the

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

predictor in the regression with the posttest measure as dependent variable (criterion).

Parametric tests. This ANCOVA of the posttest measures on the pretest measures and the Group variable did not reveal significant p values, neither for the covariate, $p = 0.9558$, nor for the Group variable, $p = 0.0576$. The absence of a significant Group effect here indicated that the Group $\times$ Time interaction was not significant. The wrapper F-value for this test was just significant, $p = 0.0474$.

Nonparametric tests. Residual rank scores were calculated. First, the pretest and posttest measures were ranked separately, ignoring Groups. Then a regression was run of the posttest ranks on the pretest ranks, again ignoring Groups, and residuals (residual rank scores) were saved. An ANOVA on such residuals corresponds to the rank analysis of covariance (RANCOVA) proposed by Quade (1967). The p value for the Group effect on the residuals (i.e., Group $\times$ Time interaction) was significant, $p = 0.0048$. The Mantel-Heanszel statistic on the differences in mean residual rank scores between the two Groups was also significant, $p = 0.0088$, as was the exact Wilcoxon two-sample test on these residual ranks, $p = 0.0057$. The permutation test option within the NPAR1WAY procedure applied to these residual ranks was also significant, $p = 0.0061$. A RANCOVA of the ranked posttest on the ranked pretest as covariate for the Group factor also returned a significant F-value, $p = 0.0031$.

Example 2: A 3 $\times$ 5 Between $\times$ Within Design with Data on an Ordinal Scale

The second example, from Shah and Madden (2004), is the powdery mildew of wheat data. This corresponds to a 3 between $\times$ 5 within factors design with three wheat cultivars (Groups) and five severity of decay (mildew) assessments (Times) and with four replications ($n = 4$ observations) of each cultivar ($N = 12$ in total). The assessments made were on a 0-to-10 ordinal (ordered categories) scale for the severity of decay. The resulting p -values for the Group $\times$ Time interaction obtained by Shah and Madden (2004) with the new rank-based methods were compared to those obtained by parametric tests, randomization tests, and two nonparametric procedures. The Beasley and Zumbo (2009) SAS/IML code was run on the data to obtain the aligned regular rank test, the aligned Friedman rank test, and the aligned Koch rank test. The ARTool was also applied on these data. The resulting p values are displayed in the first part of Table 2.

Table 2. Resulting p -values with different methods of analysis for the interaction in Examples 2, 3, and 4

3 B × 5 W		2 W × 3 W		3 B × 2 B × 2 W	
<i>p</i> -values		<i>p</i> -values		<i>p</i> -values ¹	
Parametric/randomization		Parametric/randomization		Parametric/randomization	
F-test	0.0710	F-test	0.0043	F-test	0.0692
Wrapper F-test	0.0735	Wrapper F-test	0.0046	Wrapper F-test	0.0728
Nonparametric		Nonparametric		Nonparametric	
Brunner F1_LD_F1		Brunner LD_F2		Brunner F2_LD_F1	
F-appr. small N	0.1421	F-appr. large N	< 0.0001	F-appr. large N	< 0.0001
		F-appr. small N	0.0015	F-appr. small N	0.0412
Aligned ranks		Aligned ranks		Aligned ranks	
regular	0.0427	regular	0.0099	regular	
Friedman	0.0054	Friedman	0.0065	short format	0.0753
Koch	0.0339	ARTool	0.0099	long format	0.1528
ARTool	0.0544			Friedman	0.0133
				ARTool	0.1425

¹Note: Only the triple interaction p -values are reported

Parametric tests

For the data in long format, Levene's test for homogeneity of variance between Groups was not significant, $p = 0.0669$. For the data in short format, the Mauchly sphericity test (for transformed variates) was also not significant, $p = 0.2053$. The Greenhouse-Geiser (G-G) epsilon was $\varepsilon = 0.6220$. O'Brien and Kaiser (1985) suggested that, when you have a large violation of sphericity (e.g. $\varepsilon < 0.70$) and your sample size is greater than $k + 10$ (i.e., the number of levels of the repeated measures factor + 10), then a MANOVA is more powerful; in other cases, the repeated measures design should be selected. Here, the $N = 12$ was not larger than 15 (5 time measures + 10), so we further concentrated on the repeated measures design. The p value for the F-test on the Group × Time interaction was not significant, $p = 0.0710$. The wrapper randomization test resulted in a similar p -value: 0.0735.

Nonparametric tests

Brunner's macro F1_LD_F1 also resulted in a p -value which was not significant for small samples, $p = 0.1421$. The value for large samples could not be calculated because the covariance matrix was singular and hence could not be inverted.

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

The Beasley and Zumbo (2009) procedure yielded all significant p -values, 0.0427, 0.0054, and 0.0339, respectively, for the aligned regular, Friedman, and Koch rank tests. The ARTool resulted in a non-significant value $p = 0.0544$. The correlation of the alignment with the ARTool and the alignment with the H&T formula was $r(N = 60) = 0.7811$, again indicating that the alignments were not the same.

Example 3: A Doubly Repeated Measures, 2×3 Within $\times$ Within Design, Small Sample Size

The aligned data for testing within $\times$ within, $A \times B$ interactions is given by the same H&T formula (2) as for the between $\times$ between design. In this example, I fabricated data for a 2 within $\times$ 3 within design. There were six repeated measures; one record per observation, with six columns as doubly repeated measures A and B. The means of repeated measures 1 + 2 + 3 corresponded to A level 1 and the means of the measures 4 + 5 + 6 corresponded to A level 2. The means of measures 1 + 4, 2 + 5, and 3 + 6 related to the B levels 1, 2, and 3, respectively. The results of the different tests for this example are reported in the middle part of Table 2.

Parametric tests

For this example the data, with $N = 12$ observations, were made up to yield a significant interaction, $p = 0.0043$. The randomization test wrapper resulted in an equally significant interaction value, $p = 0.0046$. The G-G ε was quite large: 0.9270.

Nonparametric tests

Brunner's macro LD_F2 is suited for doubly repeated measures. When applied on these data, both F-approximation tests, for large and small sample sizes, were quite significant interaction: $p < 0.0001$ and $p = 0.0015$, respectively.

The Beasley and Zumbo (2009) script conducted on these data confirmed the significant interaction, with p values = 0.0099 and 0.0065, respectively, for the aligned regular rank and the aligned Friedman rank tests. To achieve the Friedman rank test, for each observation, the aligned data were ranked from 1 to 6, and then a 2×3 repeated measures ANOVA was applied on these ranks. Note that Koch's ranking method is not applicable with doubly repeated measures because there are no between Groups. The ARTool returned a p value which was the same as with

the aligned regular rank test, $p = 0.0099$. The correlation of the alignment with the ARTool and the alignment with the H&T formula was perfect here, $r(N = 72) = 1$.

Example 4: A Three-Way $3 \times 2 \times 2$ Between $\times$ Between $\times$ Within Design

The fourth example was taken from Cody and Smith (1987, p. 159), a three-factor experiment with repeated measures on the last factor. They invented data for a market experiment on male and female (Sex factor) subjects who were offered three different Brands of coffee. Each brand was tasted twice (Time factor); once after breakfast, once after dinner. The preference of each brand was measured on a 10 point scale. The Shapiro Wilk test on this taste measures, over Brand, Sex, and Time combinations (in long format), was not significant, $p = 0.9465$. There were three subjects in each Brand $\times$ Sex condition combination, resulting in a total $N = 18$.

Parametric tests

There were very significant Time and Brand effects, both with $p < 0.0001$, but the triple interaction Brand $\times$ Sex $\times$ Time was not significant, $p = 0.0692$. The two-way interactions, Brand $\times$ Sex, Time $\times$ Brand and Time $\times$ Sex, were also not significant. In this example we concentrated only upon the triple interaction. The results of the different ways of analysis for the triple interaction are presented in the last part (on the right) of Table 2. The randomization test wrapper also showed a non-significant $p = 0.0728$ for this interaction.

Nonparametric tests

Brunner's macro F2_LD_F1 is suited for this design with two between factors and one repeated measures factor. When applied on these data, both F-approximation tests for the triple interaction Brand $\times$ Sex $\times$ Time, for large and small sample sizes, were significant: $p < 0.0001$ and $p = 0.0412$, respectively.

The data were aligned with the H&T schematic formula (4) for the triple interaction. The aligned data were first ranked in short format, with the two aligned repeated measures ranked separately, but over all Brand and Sex levels combinations. On these ranked aligned data, the triple interaction was not significant, $p = 0.0753$. Yet the Friedman ranking on the aligned data in short

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

format resulted in a significant triple interaction, $p = 0.0133$. Note that the Koch ranking was not applied because this is too complicated in this example.

When ranking the data aligned in long format (over Brands, Sex, and Time), as is done with the ARTool, the ANOVA yielded an F-value which was also not significant, but with a somewhat larger $p = 0.1528$. The same test on the ARTool ranked data was also not significant, $p = 0.1425$. The correlation between H&T formula aligned data and the ARTool alignment for the triple interaction was almost perfect here, $r(N = 36) = 0.9997$. The alignments were not exactly the same due to small differences in numerical precision. In the calculations of the means, only six decimals were used, whereas the ARTool uses nine.

Because of the discrepancy (in this example, not in the other examples) between the p values obtained with rankings of the aligned data in short versus long format, the ranking data were inspected and it was found that the discrepancy was due to the fact that the ranking in long format was more precise; it ranged from 1 to 36 (18 observations $\times$ 2 Times) and contained 8 ties. However, the rankings in short format twice ranged from 1 to 18 and had 10 ties and 3 doubles (equal ranks).

Discussion

Example 1: The Pretest-Posttest Design

Bonate (2000, p. 103) proposed two ways for dealing with an outlier: simply removing the outlier or applying a method to minimize the influence of an observation on parameter estimations, namely the iterative reweighted least-squares (IRWLS). Two weight functions were used for the iterations, the Huber function and the bisquare. Removing the outlier from the data in this example resulted in a p value < 0.0001 , as did both weight functions. Fagerland (2012) suggested using the Yuen-Welch t-test for trimmed means in situations with outliers. Applying this test here returned a similar p -value < 0.0001 .

Gain scores

All parametric tests for the interaction on the gain scores used in this example were not significant, apparently because of the outlier. The randomization tests only confirmed these values. Yet all parametric tests, after ranking the gain scores, were significant (< 0.01), and the nonparametric procedures also all resulted in significant p values (< 0.01).

Repeated measures

With respect to the repeated measures analyses of the interaction, the parametric test and randomization test both were not significant, but the parametric tests, after ranking the repeated measures, were also not significant (contrary to such tests after ranking the gain scores). The new rank-based methods yielded either barely well or barely not significant p -values, whereas the analyses on the aligned regular, Friedman, and Koch rank tests were clearly significant (< 0.01). The ARTool application gave a very significant p value, but the correlation between the ARTool aligned and the H&T aligned data was somewhat poor (0.7386), indicating that the ARTool alignment may not be flawless in this repeated measures design. The next smallest p -value was for the aligned Friedman ranking.

(R)ANCOVA

In the literature, there is controversy about gain scores versus repeated measures and ANCOVA (see e.g., Senn, 2006; Knapp & Schafer, 2009; Smolkowski, 2013). I do not go into this discussion here. Dimitrov and Rumrill (2003) concluded that the analysis of gain scores is okay, but that the ANCOVA should be the preferred method for analysis of pretest-posttest data. The ANCOVA (especially the wrapper version) was somewhat less affected by the outlier, which seems to support Dimitrov and Rumrill's (2003) conclusion. The RANCOVA of the ranked posttest on the ranked pretest as covariate for the Group factor and the tests on the residual ranks all gave p values < 0.01 . Dimitrov and Rumrill (2003) also noted that the attractive characteristic of residual scores is that, unlike gain score, they do not correlate with observed pretest scores. Compared with the ANCOVA model, however, they specified that an ANOVA on residuals (e.g., Quade's RANCOVA) is less powerful (p. 161).

Example 2: The 3×5 Between $\times$ Within Design

This two-way mixed 3×5 design was on an ordinal scale. With Likert-type or ordinal data, nonparametric tests are usually advised (see e.g., Kaptein et al., 2010; De Winter & Dodou, 2012), especially with small sample sizes, as in this example. Fagerland (2012) pointed to a paradox in statistical practice in high-impact medical journals, namely that the median sample size research studies published has increased manifold, while the use of nonparametric tests has increased at the expense of t -tests. It was concluded nonparametric tests should only be used with

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

small samples, because such tests in large studies may provide answers to the wrong question.

The parametric tests p -values for the interaction were not significant. All nonparametric tests in this example revealed that the interaction was significant, except for the ARTool application and the Brunner methods. The correlation between the ARTool aligned data and the H&T aligned data was slightly poor (0.7811) again, as in Example 1, showing that the ARTool does not align the data faultlessly in this design. The aligned regular rank, the aligned Friedman rank, and the aligned Koch rank tests all returned p -values < 0.01 .

Example 3: The Doubly Repeated 2×3 Within $\times$ Within Design

The data for this example were fabricated to result in a significant p -value for the interaction. All types of tests revealed this significant interaction with p -values < 0.01 . The Brunner F-approximation for large samples was even < 0.0001 . Here, the ARTool alignment and ranking were the same as the H&T alignment and ranking.

Example 4: The $3 \times 2 \times 2$ Between $\times$ Between $\times$ Within Design

This higher order design did not have a significant F-value for the triple interaction when tested parametrically or with the randomization test wrapper. Yet, when tested with the Brunner new rank-based methods or with the aligned Friedman ranking, the resulting p -values were significant. All other nonparametric tests were not significant. The Shapiro-Wilk test indicated that the distribution was quite normal and the interaction should not really be tested with nonparametric tests. This implies that significant p -values for the triple interaction in this example may be invalid.

The small discrepancy between the p values obtained after ranking in short versus long format was most probably due to the fact that the ranking in long format was more precise and less affected by ties.

Conclusion

The obtained p -values for the parametric analyses of interactions in all examples were very comparable to the values obtained by the randomization methods. In Example 1, the outlier masked the interaction with the parametrical tests as well as with randomization tests. In my opinion, the randomization tests can be seen only

as confirmation/rejection tools, not as nonparametric alternatives; it is as if they copy the parametric results.

With respect to the ARTool application, for fully between designs and fully within designs, and especially for higher order (> 2 way) such designs, this tool is indeed very convenient, as claimed by Wobbrock et al. (2011). Yet the ARTool alignment of the data for the interactions was questionable in Examples 1 and 2. Further research could perhaps clarify way the ARTool alignment of the data in these two examples was different from the H&T alignment.

The new rank-based methods can be an alternative way of looking at the data. They test treatment effects without assuming differences in location parameters (e.g., means). They test differences among distributional characteristics. One problem might be that it is not always precisely clear how population distributions differ, as noted by Serlin and Harwell (as cited in Beasley & Zumbo, 2009, p. 19). In Examples 1 and 2, these methods showed a non-significant p value for the F-approximations for small samples; this, in contrast to all (except for ARTool in Example 2) other nonparametric tests' p -values. On the other hand, in the last example, only the Friedman aligned ranks tests and this Brunner technique showed a significant triple interaction, although the data fabricated by Cody and Smith (1987) did not have a significant triple interaction with a parametric test and the distribution was quite normal. In the examples (except for Example 3) the new rank-based methods contradicted the results of other (except for the Friedman) nonparametric tests.

All the examples showed that the aligned (regular) rank test with the Higgins and Tashtoush formulae was quite sensitive to detect the interactions which should be spotted as significant. Except for pretest-posttest design, this test should be applied in factorial designs with repeated measures. In pretest-posttest designs, the RANCOVA (not Quade's ANOVA on residuals) should be the preferred way of testing. Ranking should be done on the aligned data in long format because it is more precise than separate rankings of aligned repeated measures (short format). The aligned Friedman rank tests were significant in all examples. It is advisable to report both the aligned regular rank test and the aligned Friedman rank test. They rank the data differently; the aligned regular rank test does so across observations, the Friedman test ranks the data within observations.

References

- Akritis, M. G. (1990). The rank transform method on some two factor designs. *Journal of the American Statistical Association*, 85(409), 73-78. doi: 10.1080/01621459.1990.10475308
- Akritis, M. G., & Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *Journal of the American Statistical Association*, 89(425), 336-343. doi: 10.1080/01621459.1994.10476475
- Akritis, M. G., Arnold, S. F., & Brunner, E. (1997). Nonparametric hypothesis and rank statistics for unbalanced factorial designs. *Journal of the American Statistical Association*, 92(437), 258-265. doi: 10.1080/01621459.1997.10473623
- Akritis, M. G., & Brunner, E. (1997). A unified approach to rank tests in mixed models. *Journal of Statistical Planning Inference*, 61(2), 249-277. doi: 10.1016/S0378-3758(96)00177-2
- Beasley, T. M., & Zumbo, B. D. (2003). Comparison of aligned Friedman rank and parametric methods for testing interactions in split-plot designs. *Computational Statistics & Data Analysis*, 42(4), 569-593. doi: 10.1016/S0167-9473(02)00147-0
- Beasley, T. M., & Zumbo, B. D. (2009). Aligned rank tests for interactions in split-plot designs: Distributional assumptions and stochastic homogeneity. *Journal of Modern Applied Statistical Methods*, 8(1), 16-50. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol8/iss1/4/>
- Bonate, P. L. (2000). *Analysis of pretest-posttest designs*. London, UK: Chapman & Hall/CRC.
- Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial experiments*. New York, NY: J. Wiley.
- Brunner, E., & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statistical Papers*, 42(1), 1-52. doi: 10.1007/s003620000039
- Cassell, D. L. (2002). A randomization-test wrapper for SAS® PROCs. *Proceedings of the Twenty-seventh Annual SAS users Group International Conference*. Cary, NC: SAS Institute Inc. Retrieved from <http://www2.sas.com/proceedings/sugi27/Proceed27.pdf>
- Cody, R. P., & Smith, J. K. (1987). *Applied statistics and the SAS programming language* (2nd ed.). New York, NY: North-Holland.

- Conover, W. J., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communication in Statistics – Theory and Methods*, 5(14), 1349-1368. doi: [10.1080/03610927608827447](https://doi.org/10.1080/03610927608827447)
- De Winter, J. C. F., & Dodou, D. (2010). Five-point Likert items: t test versus Mann-Whitney-Wilcoxon. *Practical Assessment, Research & Evaluation*, 15(11). Retrieved from <http://pareonline.net/getvn.asp?v=15&n=11>
- Dimitrov, D. M., & Rumrill, P. D., Jr. (2003). Pretest-posttest designs and measurement of change. *Work*, 20(2), 159-165.
- Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies – a paradox of statistical practice? *BMC Medical Research Methodology*, 12(78). doi: [10.1186/1471-2288-12-78](https://doi.org/10.1186/1471-2288-12-78)
- Gao, X., & Alvo, M. (2005a). A nonparametric test for interaction in two-way layout. *The Canadian Journal of Statistics*, 33(4), 529-543. doi: [10.1002/cjs.5550330405](https://doi.org/10.1002/cjs.5550330405)
- Gao, X., & Alvo, M. (2005b). A unified nonparametric approach for unbalanced factorial designs. *Journal of the American Statistical Association*, 100(471), 926-941. doi: [10.1198/016214505000000042](https://doi.org/10.1198/016214505000000042)
- Hettmansperger, T. F. (1984). *Statistical inference based on ranks*. New York: Wiley.
- Higgins, J. J., & Tashtoush, S. (1994). An aligned rank transform test for interaction. *Nonlinear World*, 1(2), 201-211.
- Hodges, J. L., Jr., & Lehmann, E. L. (1962). Rank methods for combination of independent experiments in the analysis of variance. *Annals of Mathematical Statistics*, 33(2), 482-497. doi: [10.1214/aoms/1177704575](https://doi.org/10.1214/aoms/1177704575)
- Kaptein, M., Nass, C., & Markopoulos, P. (2010). Powerful and consistent analysis of Likert-type rating scales. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.
- Kelley, D. L., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the f statistic in analysis of variance. *Journal of Statistical Computation and Simulation*, 58(4), 343-359. doi: [10.1080/00949659708811839](https://doi.org/10.1080/00949659708811839)
- Knapp, T. R., & Schafer, W. D. (2009). From gain score t to ANCOVA F (and vice versa). *Practical Assessment, Research & Evaluation*, 14(6). Retrieved from <http://pareonline.net/getvn.asp?v=14&n=6>
- Leys, C., & Schumann, S. (2010). A nonparametric method to analyze interactions: The adjusted rank transform test. *Journal of Experimental Social Psychology*, 46(4), 684-688. doi: [10.1016/j.jesp.2010.02.007](https://doi.org/10.1016/j.jesp.2010.02.007)

NONPARAMETRIC RANK TESTS FOR INTERACTIONS

McSweeney, M. (1967). An empirical study of two proposed nonparametric tests for main effects and interaction (Doctoral dissertation). Retrieved from Dissertation abstracts international. (1968-16394-001)

Noguchi, K., Gel, Y. R., Brunner, E., & Konietzschke, F. (2012). nparLD: an R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, 50(12). doi: [10.18637/jss.v050.i12](https://doi.org/10.18637/jss.v050.i12)

O'Brien, R. G., & Kaiser, M. K. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, 97(2), 316-333. doi: [10.1037/0033-2909.97.2.316](https://doi.org/10.1037/0033-2909.97.2.316)

Peterson, K. (2002). Six modifications of the aligned rank transform test for interaction. *Journal of Modern Applied Statistical Methods*, 1(1), 100-109. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol1/iss1/13/>

Puri, M. L., & Sen, P. K. (1985). *Nonparametric methods in general linear models*. New York, NY: Wiley.

Quade, D. (1967). Rank analysis of covariance. *Journal of the American Statistical Association*, 62(320), 1187-1200. doi: [10.1080/01621459.1967.10500925](https://doi.org/10.1080/01621459.1967.10500925)

R Core Team. (2012). *R: a language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Richter, S. J., & Payton, M. E. (1999). Nearly exact tests in factorial experiments using the aligned rank transform. *Journal of Applied Statistics*, 26(2), 203-217. doi: [10.1080/02664769922548](https://doi.org/10.1080/02664769922548)

Rodriguez, O. J. C., Álvarez, G. J., & Ramirez, R. J. (2009). Análisis no paramétrico de la interacción de dos factores mediante el contraste de rangos alineados [An aligned rank test for a nonparametric analysis of the two-way interaction]. *Psicothema*, 21(1), 152-158.

Salazar-Álvarez, M., Tercero-Gómez, V. G., Temblador-Pérez, M., & Conover, W. J. (2014). Nonparametric analysis of interactions: a review and gap analysis. *Proceedings of the 2014 Industrial & Systems Engineering Research Conference*. Montréal, Canada.

SAS Institute Inc. (2008). The SAS System, version 9.2 [Computer software]. Cary, NC: SAS Institute Inc.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, 60(1), 91-126. doi: [10.3102/00346543060001091](https://doi.org/10.3102/00346543060001091)

- Sen, P. K. (1968). On a class of aligned rank order tests in two-way layouts. *Annals of Mathematical Statistics*, 39(4), 1115-1124. Retrieved from <http://www.jstor.org/stable/2239679>
- Senn, S. (2006). Change from baseline and analysis of covariance revisited. *Statistics in Medicine*, 25(24), 4334-4344. doi: 10.1002/sim.2682
- Shah, D. A., & Madden, L. V. (2004). Nonparametric analysis of ordinal data in designed factorial experiments. *Phytopathology*, 94(1), 33-43. doi: 10.1094/PHYTO.2004.94.1.33
- Smolkowski, K. (2013). *Gain score analysis* [Unpublished report]. Retrieved from http://homes.ori.org/~keiths/Tips/Stats_GainScores.html
- Toothaker, L. E., & Newman, D. (1994). Nonparametric competitors to the two-way ANOVA. *Journal of Educational & Behavioral Statistics*, 19(3), 237-273. doi: 10.3102/10769986019003237
- Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The Aligned Rank Transform for nonparametric factorial analyses using only ANOVA procedure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. New York, NY: ACM Press.

Does One Size Fit All? A Case for Context-Driven Null Hypothesis Statistical Testing

Grayson L. Baird
Rhode Island Hospital
Providence, RI

Lisa L. Harlow
University of Rhode Island
Kingston, RI

Rodgers (2010a) asserted that the practice of null hypothesis statistical testing (NHST) follows a mechanistic and rule-based epistemology. This concern is addressed using historical and modern sources as evidence for NHST as a dynamic, context-driven framework for empowering researchers in scientific inquiry.

Keywords: Null hypothesis statistical testing, NHST, context-driven

---

Introduction

Rodgers (2010a; 2010b) brought to light many important issues pertaining to what he called a “quiet revolution” (p. 2) concerning statistics in practice. Rodgers noted the practice of null hypothesis statistical testing (NHST) follows a “mechanistic” (p. 10) and “rule-based” (p. 1) epistemology. The intent of this article is to elaborate on this idea and to consider how the current NHST framework is applied in a somewhat rigid and prescriptive fashion. Specifically, the automatic and what Cohen (1994) called “ritual” (p. 997) practices of NHST is examined relative to what was suggested in original sources by foundational theorists (e.g., Fisher, 1926; 1928; 1935; 1973; Neyman & Pearson, 1933a; 1933b; Yule & Kendall, 1950). In addition, original and contemporary sources are provided as evidence for NHST as a dynamic, context-driven framework for empowering researchers in scientific inquiry. Although ritualism may be pervasive throughout many aspects of NHST, the scope of this paper is limited to considering only the selection of the critical value and the value of the null hypothesis.

Grayson L. Baird is a research statistician with the Lifespan Biostatistics Core at Rhode Island Hospital. Email him at: grayson_baird@brown.edu.

The Critical Value (α) and Level of Significance

Rodgers (2010a) described the practice of NHST as a set of procedures, applied mechanistically. Researchers today often collect data and test their hypotheses by deriving test statistics and corresponding p -values, from which statistical significance of results is ascribed if the derived p -value is less than a fixed threshold, conventionally 0.05 (for a concise history of 0.05 level of significance, see Cowles & Davis, 1982). This fixed threshold is used as a conventional cutoff value for determining if a result is statistically significant, above random variation, assuming the null. In practice, 0.05 (or alternatively, sometimes 0.01 and 0.001, see Skipper, Guenther, & Nass, 1967) is almost the universal definition of significance regardless of the subject area, the nature and size of the sample, the quality of the measurement, the quality and nature of the design, the hypothesized and actual effect size, or the research question itself.

Although the practice of using 0.05 is pervasive, a great deal of criticism towards NHST results from the use of an arbitrary and traditional cutoff value to determine significance (see Mudge, Baker, Edge, & Houlahan, 2012). For instance, early on, Selvin (1958) noted “reciting the magic phrase ‘significant at the 0.01 level’ is often a substitute for hard thinking about the quality of one’s data” (p. 86). Ironically, this ritualistic practice of determining significance does not appear to be in accordance with testing espoused by either Neyman and Pearson or Fisher. Specifically, when discussing errors of the first and second kind (i.e., Type I error (PI), rejecting a null hypothesis that should be retained, and Type II error (PII), holding onto a null hypothesis that should be rejected, respectively), Neyman and Pearson (1933a) noted:

These two sources of error can rarely be eliminated completely; in some cases it will be more important to avoid the first, in others the second. We are reminded of the old problem considered by Laplace of the number of votes in a court of judges that should be needed to convict a prisoner. Is it more serious to convict an innocent man or to acquit a guilty?... From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimized. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator. (p. 296)

Neyman and Pearson (1933b) also noted: “we attempt to adjust the balance between the risks PI and PII, to meet the type of problem before us” (p. 497). Here,

DOES ONE SIZE FIT ALL? A CASE FOR CONTEXT-DRIVEN NHST

Neyman and Pearson described a system whereby the researcher plays an active role in evaluating significance, in the context of minimizing and thus balancing errors of the first and second kind, which are inversely related to each other, relative to the conditions of the study at hand. Therefore, significance level is not an arbitrary and universal value, but rather a value that achieves a meaningful and appropriate balance of Type I versus Type II errors, determined by the researcher with the specific conditions of the study in mind. Neyman and Pearson (1933b) stressed the influence of context for deciding if a small or large critical value is warranted.

However, some of Fisher's writings may be viewed as promoting a fixed level of significance. For instance, Fisher (1928) noted:

[Regarding] the value for which $P = .05$, or 1 in 20... it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. (p. 45)

In this statement, Fisher appears to have advocated a significance level of 0.05. However, also around this time, Fisher (1926) wrote:

If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent point), or one in a hundred (the 1 per cent point). Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance. (p. 504)

This statement reveals that a significance level of 0.05 is viewed by Fisher as a "low" standard and other levels of significance may be used. Of the two aforementioned statements, attention should be drawn to Fisher's use of the words "convenient" and "prefers" when he described choosing a level of significance. Here, Cochran (1976) suggested that Fisher appeared to be promoting a significance level of 0.05 based on preference but not advocating 0.05 as an exclusive level of significance. Also apparent in Fisher's aforementioned statement is that his confidence in experimental results rested with the quality of the design. Further evidence of Fisher's reluctance to assign an official level of significance

but instead consider significance in light of the conditions of the research can be seen in the following statement some years later:

...the attempts that have been made to explain the cogency of tests of significance in scientific research, by reference to supposed frequencies of possible statements, based on them, being right or wrong, thus seem to miss the essential nature of such tests. A [scientist] who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. . . . However, the calculation is absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, [they] reject hypotheses; [the scientist] rather gives [their] mind to each particular case in the light of ... evidence and [one's] ideas. (1973, p. 45)

Fisher (1973) suggested a level of significance that may reliably indicate statistical significance over the long run, although he was quick to condition this statement by noting that a universal or fixed level of significance used in all situations would not make sense. He noted, "In choosing the grounds upon which a general hypothesis should be rejected, personal judgment may and should properly be exercised" (p. 50).

This is evidence neither Neyman and Pearson nor Fisher advocated any universal or canonical level of significance, but rather entrusted that researchers would define a level of significance that was relevant to their field of research and appropriate for the conditions of the study. Specifically, it is possible to see that Neyman and Pearson advocated testing as a dynamic procedure where the researcher actively engages in evaluating "what is significant" by balancing the costs of committing Type I versus Type II errors relative to the context of the research. It is also possible to see that Fisher was advocating testing also as a dynamic procedure, where the researcher actively engages in evaluating what is significant by considering the conditions of the particular study and the nature of the research question.

It is difficult to imagine either Neyman and Pearson or Fisher as supporters of mechanistic thinking in general. Pearson (1955) noted that "from the start we shared Professor Fisher's view that in scientific inquiry, a statistical test is 'a means of learning'" (p. 206). Neyman asserted "for a satisfactory performance of a statistician's duty... it is necessary that [they] fully understand the circumstances of experiments, whatever their nature, to which statistical methods are applied"

DOES ONE SIZE FIT ALL? A CASE FOR CONTEXT-DRIVEN NHST

(Reid, 1982, p. 183). Fisher specifically noted that “tests of significance are used as an aid to judgment, and should not be confused with automatic acceptance tests” (Fisher, 1928). None of the aforementioned statements presented suggest that any of these theorists intended for or engendered a ritualistic application of statistics.

Fisher and Neyman and Pearson advocated two separate frameworks. For example, Neyman-Pearson theory advocated setting the critical value (α) a priori of analysis, whereas Fisher advocated reporting significance level after analysis. Fisher’s framework tested the null only, whereas the Neyman-Pearson framework tested two or more hypotheses, thus allowing errors of the first and second kind to be controlled for and power to be estimated. A more in-depth review of these and other differences can be found in Gigerenzer (2004). Although a greater elaboration of these differences is beyond the scope of this review, it is important to establish that neither framework appears to be advocating an arbitrary and universal threshold of significance, such as 0.05.

Fisher’s and Neyman and Pearson’s treatments of significance as a contextual judgment appears to be in agreement with other original theorists. For instance, Yule and Kendall (1950) noted:

In the examples we have given...our judgment whether P was small enough to justify us in suspecting a significant difference...has been more or less intuitive. Most people would agree...that a probability of only 0.0001 is so small that the evidence is very much in favour of the supposition that the dice were biased...Suppose we had obtained $P = 0.1$ ...Where, if anywhere, can we draw the line? The odds against the observed event which influence a decision one way or the other depend to some extent on the caution of the investigator. Some people (not necessarily statisticians) would regard odds of ten to one as sufficient. Others would be more conservative and reserve judgment until the odds were much greater. It is a matter of personal taste. (p. 471)

This discussion of significance by Yule and Kendall appeared to have advocated a way of determining significance based on a researcher’s intuition, caution, scientific background, and “personal taste.” It should be noted that they went on to mention that there are two values of P , 0.05 and 0.01, which are used widely to provide a “rough line of demarcation” for level of significance (p. 472). Yule and Kendall do not appear to have promoted a strict level of significance; rather, they appear to have advocated a significance level based on contextual considerations and later mentioning 0.05 and 0.01 as rough thresholds commonly

used. Thus, significance level based on contextual considerations appears to be an established tradition early on in the field of statistics.

The convention of statistical significance being achieved if and only if p is less than an arbitrary and recognized cutoff is perhaps the most illustrative instance of rather rote thinking in the current practice of applied statistics in psychology. The current NHST framework, in practice, allows arbitrary, traditional, and preordained cut off values to determine the significance of the results rather than allowing the significance of the results to be determined by researchers with the conditions of the study and the research question in mind. Thus, pervasively defining significance at 0.05 has led the process of inference away from a scientific basis, as noted by Morrison and Henkel (1969):

If, indeed, .05 (or any other level) is ‘sacred’...then what do we do in sociology surely is much more akin to religion than science and we might as well forget empirical work and get on with the development of more rituals. (p. 137)

Fortunately, there is support for NHST, as a framework, which empowers researchers to evaluate the significance of their results relative to the context of their research. Aguinis et al. (2010) asserted that conventional cutoffs ignore the relative seriousness of committing a Type I versus II error for a given study. For instance, researchers studying the possible effects of a new drug could be committing a Type I error if the drug was found to appear effective although it was later found to have very serious side effects that reduced the benefit of the treatment and potentially endangered the patients. Alternatively, researchers may commit a Type II error when testing a drug with little or no side effects if the power in their study is small (due, for example, to a small sample size and/or a small effect), which could have serious consequences by potentially removing a viable treatment from consideration by patients needing new options. Thus, in some research contexts, failing to reject the null when the null is false may be more serious than rejecting the null when the null is true (or vice versa). As a consequence, the widely used convention of maintaining arbitrary cutoffs disallows the researcher to appreciate and control for the relative seriousness of committing either Type I or II errors in a given research situation.

Aguinis et al. (2010) therefore proposed a “customer-centric” (p. 517) approach to science, where the customer (i.e., the researcher) controls the probability of committing a Type I and Type II error based on the relative seriousness of committing these errors and given the nature of the research and research question. Thus alpha is chosen by the researcher based on the context of

the research and predicated on the researcher's preference and rationale behind the relative seriousness of committing a Type I versus II error.

A similar proposal by Baker and Mudge (2012; also see Mudge et al., 2012) called for researchers to explicitly consider the relative costs of Type I and II errors when determining a value of alpha. Baker and Mudge pointed out that when sample variability is high and/or the sample size is low, the habitual use of designating alpha at 0.05 leads researchers to unrealistically test for real effects and, as a consequence, increase the rate of Type II errors (false negatives). They advocated using an optimal alpha value that takes into account the relative costs of committing Type I and II errors using power analysis (e.g., Cohen, 1988); however, instead of determining a needed value for power, effect size, or sample size, alpha is being determined. Thus, researchers can calculate an optimal alpha value by specifying a meaningful (a priori) critical effect size, sample size, and different values of power. In practice, once the observed sample size and a meaningful effect size have been specified, if a Type II error is more serious than a Type I, then simply increase the value of power which will decrease the probability of a Type II error, thereby increasing the probability of a Type I error. Conversely, if a Type I error is more serious than a Type II, then simply decrease the value of power which will in turn decrease the probability of a Type I error but will increase the probability of a Type II error.

However, Baker and Mudge (2012) held that, for most studies, alpha should be a value that minimizes the overall probability or cost of making a mistake; thus the selected alpha value should minimize the combined probabilities of a Type I and II error. Specifically, they noted that “If we consider minimising the chances of errors to be the goal for good decision-making, we can choose an optimal decision-making threshold (optimal α level) that minimises the average of α and β (Type I and Type II errors) at the smallest potentially meaningful effect size” (p. 30). They also asserted that researchers should report the sample size, observed variability of the data, exact p -values, specified power value and effect size used in determining each optimal alpha value so that other researchers can re-evaluate results using different optimal alpha values based on their own notions of relative cost of Type I and II error and critical effect sizes. Thus, instead of convention, the context of the study (e.g., relative seriousness of Type I and II error, or the goal of reducing both errors optimally, the variability of the data, the observed sample size, the a priori desired or hypothesized effect size, and power) must be considered when setting an alpha value.

Cascio and Zedeck (1983) proposed directly assessing the “apparent relative seriousness” (ARS) of Type I and II errors with the following equation:

$$ARS = \left[\frac{P(H_1) + \beta}{1 - P(H_1) + \alpha} \right] \quad (1)$$

where $P(H_1)$ is the probability that the null is false, $1 - P(H_1)$ is the probability that the null is true, β is the probability of a Type II error, and α is the probability of a Type I error. In essence, if alpha is held constant at 0.05 and power 0.80 then, as the probability of the null being false increases, the relative importance of alpha increases dramatically; thus when $P(H_1) = 0.1, 0.2, 0.4, 0.5$, or 0.7 , the alpha value is 0.44, 1, 2.7, 4, or 9.3 times more serious than β . The ARS equation allows researchers to directly assess relative seriousness of error with aspects of context within their study.

Using this framework, Murphy and Myers (2004) proposed operationalizing an appropriate alpha value based on the same aspects of research context. Therefore, instead of assessing ARS based on an arbitrary alpha value, they proposed determining a specific alpha value using a researcher's desired relative seriousness (DRS) value of committing a Type I versus Type II error (desired ARS value), given:

$$\alpha_{\text{desired}} = \left[\frac{P(H_1)\beta}{1 - P(H_1)} \right] \cdot \left[\frac{1}{\text{DRS}} \right]. \quad (2)$$

Alpha is based on the context of the research, where the balance of Type I over Type II error is specified by the researcher. Moreover, the researcher's confidence in the alternative being true along with the researcher's notion that a rejection may be false is realized mathematically. Confidence in the alternative being true and the probability of Type II error may be due to the quality of the sample, the quality and control of the design, researcher's experience, previous research, etc. The benefit here is this approach produces an alpha level that fits the needs of the researcher relative to the conditions of the study (to the degree the conditions of the study can be translated into those parameters) and can be justified a priori.

Both the historical and contemporary authors mentioned revealed a need for determining an alpha value appropriate for the context of the research instead of using conventional and universal cutoff values. However, alpha (or level of significance) is not the only value thoughtlessly selected in the application of NHST.

The Value of Null

Another common criticism of NHST is that the hypothesis being tested is often limited to the hypothesis of no effect, often called the nil hypothesis (see [Cohen, 1994](#)). There are several research situations where testing the nil is appropriate given the research question. Specifically, if one is interested in examining any effect or difference above zero only, then the nil is a logical hypothesis of comparison (e.g., effect of experimental manipulation between two randomly assigned groups). Although this type of research question may be seen as overly simplistic to many researchers, testing the nil can nevertheless legitimately address the question of interest. While testing the nil hypothesis may be statistically sound, however, the habitual practice of testing only the nil in all research contexts, as a default value rather than a null value of interest, is another illustration of ritualistic practice in the application of NHST.

The Null as a Value

Although nil hypothesis testing is often used, it is not the only hypothesis available to researchers within the NHST framework. Originally, Fisher proposed the null hypothesis as the hypothesis of interest in which we were trying to disprove; it is the hypothesis we are trying to nullify (see [Bakan, 1966](#); [Cohen, 1994](#); [Gigerenzer, 2004](#)). Thus, the hypothesis to be nullified can refer to any null value, including but not limited to the nil. Specifically, Fisher (1935) asserted “we may, however, choose any null hypothesis we please, provided it is exact” (p. 20). This is perhaps best illustrated in his infamous Lady tasting tea problem: the Lady asserted that she could discriminate between cups of tea where the milk was infused either before or after the tea was poured. Her claim is tested with 8 cups of tea, 4 containing tea with milk infused prior to pouring and 4 after. The Lady is presented with the cups in random order and is blinded to their preparation. Fisher noted that the null could be either that the Lady has no sensory discrimination in detecting how tea was prepared regarding milk or that she has perfect sensory discrimination ([Fisher, 1935, p. 13](#)). Thus, the null could be 0.5 or 1.0, revealing that the null need not be the nil (i.e., 0.5).

The Neyman-Pearson (see [Neyman, 1950; 1957](#)) hypothesis testing framework specifically required researchers to designate the values of the null and an alternative (H_1 and H_2), where the null is preferably called the “hypothesis tested” and “it is immaterial which of the two alternatives H_1 and H_2 is labeled the hypothesis tested” ([Neyman, 1950, p. 259](#)). To illustrate this point, Neyman considered two hypotheses in regards to Fisher’s Lady tasting tea problem. He

noted of (a) $p \neq 1/2$ and (b) $p = 1/2$ that one of these hypotheses will be “the hypothesis tested” and the other “the alternative hypothesis” (Neyman, 1950, p. 273). Neyman went on to say that which claim will be regarded as the hypothesis tested and which the alternative depends on the situation and the balance of errors of the first and second kind: if we were the Lady, we would want the hypothesis tested to be (a), as the more important error to avoid is having her claim refused (avoid rejecting (a) if (a) were true); if we were the jury, we would want (b), given that the more important error to avoid is the granting of an unjustified claim (avoid rejecting (b) if (b) were true). Here, context plays into which hypothesis is the “null” in concert with balancing errors of the first and second kind. In another example, Neyman (1942) provided general guidance for selecting the hypothesis to be tested; he noted that the null hypothesis should be the hypothesis whereby the errors of the first kind are of greater importance relative to errors of the second kind. In this example, he specifically chose a non-nil hypothesis (i.e., “the actual toxicity of the drug does exceed the prescribed safety limit”) given the relative importance of a Type I error (p. 304).

The two examples above concerning the Lady tasting tea experiment reveal that although Fisher and Neyman and Pearson explicitly promoted two different frameworks, neither advocated that the null always be defined as the nil. Indeed, as illustrated by both Fisher and Neyman, in theory and application, the null can be defined as any value; instead of the nil, or a value of zero, being the standard, it is just one possible hypothesis to test within the greater NHST framework (see Murphy & Myors, 1999).

Apart from reducing the involvement of researchers in the decision process, the default use of the nil as the null hypothesis can also limit application and theory. For instance, Serlin (1987) asserted that use of the nil hypothesis provides weak evidence for many theories given that it is often believed a priori that populations do in fact differ at least somewhat. In application, always testing the nil can be problematic because, most often, samples differ from each other to some degree, regardless if they come from the same population or different populations, due to sampling error alone. Meehl (1990) went so far as to call the use of nil hypothesis testing a “weak use” (p. 116) of a significance test, because the nil is (literally) always false. In addition, nil hypothesis testing is limited in detecting only if a difference or relationship exists, above zero, without regard to magnitude (Murphy & Myors, 2004). Finally, by only testing a single null value, statistical significance can be achieved by simply increasing the sample to a sufficient size (Serlin & Lapsley, 1985).

DOES ONE SIZE FIT ALL? A CASE FOR CONTEXT-DRIVEN NHST

Fortunately, these issues associated with exclusively defining the null as the nil are largely unnecessary. As mentioned, nil hypothesis testing can provide weak evidence for theories in which differences between populations or relationships between variables are anticipated or known to exist. Hodges and Lehmann (1954) noted that “when we formulate the hypothesis that the sex ratio is the same in two populations, we do not really believe that it could be exactly the same, and would only wish to reject equality if they are sufficiently different” (p. 261). One way to test for these “sufficient differences” lies in testing some value for the null other than zero. Murphy and Myers (1999) advocated an alternative to nil hypothesis testing which they termed “minimum-effect” testing. This framework is predicated on testing against a “negligibly small or trivial” effect, rather than testing for zero. Thus, depending on the context of the study, minimum effects testing can test more realistic hypotheses, rather than the “straw man” nil (Serlin & Lapsley, 1985, p. 74), which may be untenable in many research situations.

Another benefit of minimum effects testing is that it allows researchers to test both the presence of an effect and the magnitude of said effect by creating an upper and lower bound; thus, a range of null values can be tested instead of a specific value only. A minimum effect null is no longer a point hypothesis but rather a range between the minimum effect specified and the nil. Thus, if we set a null to 3% of variance accounted for and we reject this null, then we are more confident that a real effect exists because we are no longer testing a null of 0% variance accounted for. Moreover, by testing a non-nil null, when we do reject the null, we now have some information about the magnitude of said effect (e.g., the effect is above 3% variance accounted for). The benefits of using a minimum effect are apparent; however, the drawback of using a minimum effect is it increases the risk of committing a Type II error.

Although Murphy and Myers (1999) admitted that establishing a suitable minimum-effect value may be difficult initially, the benefits of such testing could greatly increase the meaningfulness of results. Thus they advocated a system whereby the hypothesis being tested is not determined for the researcher by convention, but rather the researcher determines a hypothesis relevant for the given research question and relative to the conditions of the study (e.g., a priori desired or hypothesized effect size, confirmation vs. exploratory study, theory concerning the population(s) being tested, etc.).

Use of the non-nil null also should not be applied in a rote manner. As Knapp and Sawilowsky (2001) warned, some effects are inherently small; thus, by using an arbitrary non-nil null, the chances of these (albeit) small effects being missed are increased, if not certain, depending on the non-nil value. Therefore, the value

of any null, nil or not, must be guided by context. As a consequence, this framework empowers researchers to operationalize their research questions by evaluating and designating a value of sufficient difference or relationship (minimum effect) germane to and appropriate for the area of focus.

The Null as an Interval

Originally, Hodges and Lehmann (1954) proposed testing “sufficient differences or relationships” by using a range of possible values for the null hypothesis rather than testing a single null value. Later, Meehl (1990) proposed what he called a “strong use of hypothesis test” whereby the null is a specific value a researcher asserts as their theory, and therefore as the null they are testing against their assertion (p. 79). Serlin and Lapsley’s (1985) framework advocated testing one’s own theory as the null, along with using what they call a “good-enough belt” around a “complex null hypothesis” (p. 79). Instead of testing a nil hypothesis exclusively, they recommended testing a null value that represents one’s theory (which could include the nil) and has a belt or width (denoted as Δ) around the value of the chosen null value. For example, instead of testing a null value against one’s hypothesized value, researchers instead designated their hypothesized values as the null, and use good-enough belts to test a range of possible null values (e.g., 2.5 ± 0.5); thus one can think of good-enough belts as a type of confidence interval for the null value (see Serlin, 1987). Serlin and Lapsley (1985) noted that, by using good-enough belts, the imprecision of estimating the population is reduced because a range is being tested instead of a single all-or-nothing value. Moreover, they noted that instead of simply testing a direction, researchers are testing the magnitude of the change in direction.

A major criticism of the NHST is that the null can almost always be rejected when the sample size is sufficiently large. This problem, sometimes referred to as “infinite precision” (Serlin & Lapsley, 1985, p. 74), is a function of infinite (or very large) sample size whereby natural differences between populations can be detected even if they are not meaningful (Serlin & Lapsley, 1985). Conversely, by testing a range of possible null values, the almost inevitable rejection of the null due to increasing sample size is reduced. Serlin and Lapsley (1985) noted that the value of Δ must be chosen by the researcher a priori and “reflects the state of the art or the error in the best ‘known experimental technique’ in the field” (p. 79). The framework proposed by Serlin and Lapsley empowers researchers to determine a range of meaningful null values instead of mechanistically testing a single all-or-nothing value that is more easily rejected with a large enough sample. Thus they

advocated a framework where the researcher, not ritual, decides the hypotheses of interest and where large samples do not automatically guarantee significance.

In summary, the inherent limitations associated with testing the nil hypothesis without ample consideration of a desired effect are largely unnecessary given the context-based alternative frameworks mentioned (although some may consider these to be alternatives to NHST itself; see [Denis, 2003](#)). Specifically, Murphy and Myors (1999) advocated a framework that empowers the researcher to evaluate the significance of hypotheses by determining a (minimum effect) null value that is meaningful to the researcher and appropriate for the context of the research. What is more, Serlin and Lapsley (1985) advocated a framework that empowers researchers to both specify a hypothesis of interest (including but not limited to the nil) while also determining a range or interval of possible values (a good-enough belt) where the null may still hold. Neither framework allows the researcher to blindly test a nil hypothesis by default (the dangers of which are clearly illustrated by [Sawilowsky, 2003](#)). These frameworks therefore empower researchers to specify their hypotheses in concert with the context of their research areas and questions.

Discussion

Many have observed that the current application of NHST is ritualistic (see [Cohen, 1994](#)) and mechanistic ([Rodgers, 2010a; 2010b](#)). Gigerenzer (2004) even labeled this phenomenon as “the null ritual” (p. 33). Indeed, a ritualistic approach to NHST, where the null hypothesis value and critical value are predetermined by convention, may actually impede researchers from testing the hypotheses appropriate for their particular research questions. In addition, rote selection of the nil and critical values may induce researchers to inadvertently ignore many important conditions of their study, such as the hypothesized effect size and the relative seriousness of Type I versus Type II error. As a consequence, the null ritual, not the researcher, ends up determining the significance of hypotheses and even the hypotheses themselves without regard to the context of the research. If used in this fashion, the application of NHST is indeed in danger of becoming a rite or ceremonial practice, much akin to those of the cargo cults where the deliverance of a p -value smaller than 0.05 is tantamount to a cargo box (see [Feynman, 1985](#)).

Although NHST may often be applied in practice without regard to context, there is little evidence that hypothesis testing was ever intended to be used in this fashion by original theorists. Neyman and Pearson, Fisher, Yates, and Kendall all wrote about determining significance relative to the judgment of the researcher in

concert with the context of the research itself; none appear to have advocated for the definition of the null as the nil hypothesis exclusively. In addition, contemporary authors reviewed here offer innovative ways of conceptualizing the application of NHST to better suit the context of research while breaking away from habitually testing the point nil hypothesis. By implementing the concepts from these sources, both traditional and contemporary, researchers are engaged in what could be described as “context-driven NHST” or CD-NHST. Instead of being driven by convention, which may or may not have much relevance, CD-NHST places the researcher in the driver’s seat of inference. In so doing, CD-NHST is in part responding to the changes in quantitative thinking and training called for by Rodgers (2010a; 2010b) and others (e.g., Cumming, 2012; Harlow, Mulaik & Steiger, 1997; Kline, 2011). Rodgers (2010a) noted:

The treatment of the null and alternative hypotheses, of Type I and Type II errors, and of power needs to change to accommodate the focus on the researcher’s model, rather than the null (nil) hypothesis. (p. 10)

CD-NHST not only addresses the issues brought up by Rodgers (2010a), but a happy by-product of CD-NHST is that, as a general framework, it inherently promotes replication and meta-analysis. Because CD-NHST requires more thought and detail, studies using CD-NHST could therefore yield an abundance of data for replication and meta-analytic studies. Specifically, with thoughtful and specific critical values, null values, and null ranges based on justified contextual reasons and all being reported, researchers can have access to a wealth of data to perform well-informed replications and meta-analyses. More importantly, CD-NHST as a framework relies on designating values from previous studies, thereby relying, to some degree, on replication itself.

Although the bulk of this discussion emphasizes empowering researchers by placing the selection of critical values and null hypothesis value(s) into the hands of researchers rather than being determined by common practice, this viewpoint is not without controversy. As noted by Cortina and Landis (2010), by having alpha set by externally determined criteria, corroboration between the hypothesis and data is compelling because the evidence is determined independently of the researcher; one may assume this thinking also extends to the selection of the null hypothesis value(s) as well. Conversely, Hubbard and Ryan (2000) asserted that conventional cutoffs only provide an illusion of objectivity that “makes life tidier” rather than requiring researcher’s to use subjective judgment. Indeed, although setting alpha to a default value may be objective, one should always remember inference remains

subject to the conditions of the study. Both points presented by Cortina and Landis (2010) and Hubbard and Ryan (2000) are important; thus a delicate balance must be struck between researchers evaluating their hypothesis and remaining objective in their evaluation. By empowering researchers to make context-driven decisions regarding the application of NHST, we at the same time risk inviting a certain level of subjectivity into the analysis.

One possible solution in balancing active evaluation and biased subjectivity would be to encourage researchers to establish critical values, null values, and null ranges a priori of data analysis or even data collection. This would allow researchers to participate in determining their hypotheses and the significance of said hypotheses, without the data and results influencing these decisions. A second way to encourage researchers to engage in context-driven NHST without biasing their results could be achieved by having researchers justify specifically why they are using a particular critical value or null hypothesis value (based on previous research, theory, etc.).

A third step would be to encourage researchers to report as much detail as possible in their articles. Specifically, by researchers reporting specific p -values (McGrath, 2011), confidence intervals (Cumming, 2012), effect sizes (Grissom & Kim, 2012), and power analyses (e.g., Cohen, 1988; and see Denis, 2003), readers can form their own conclusions from a given study. Beale (1972) asserted that “The p level is for the reader's use, and [the reader] alone should be the one who decides whether the p level reported is significant” (p. 1080). Reporting specific p -values has also been proposed by contemporary authors (see Aguinis et al. 2010; Baker & Mudge, 2012). Careful consideration must be used here in distinguishing the utility of alpha and p -values for CD-NHST. It is essential that authors establish an appropriate (and hopefully context-driven) alpha value which allows the authors to evaluate and conclude if a given result occurs above random variation, assuming the null; reporting p -values allows readers to evaluate the results for themselves, though this practice does not remove the real need for authors to establish a justifiable alpha value (Knapp & Sawilowsky, 2001). In addition, confidence intervals can play a special role in reporting as they contain information concerning estimation with inference, which exceeds the utility of a p -value alone. In summary, these three suggestions can help to promote researchers in engaging in context-driven NHST while also attempting to minimize the bias inherent in researchers’ decisions.

Given the aforementioned arguments, it is not difficult to at least question the wisdom of a “one-size-fits-all” approach when using NHST. However, what exactly is CD-NHST in application? Current researchers and statisticians (e.g.,

Aguinis et al, 2010; Cohen, 1994; Cumming, 2012; Mudge et al., 2012; Murphy & Myers, 2004) have aptly decried the perpetuation of the exclusive and, admittedly, somewhat mindless use of the 0.05 critical or a nil difference of zero when making statistical inferences and original theorists (e.g., Fisher, 1926; 1935; 1973; Neyman & Pearson, 1933a; 1933b; Neyman, 1950; Yule & Kendall, 1950) never seemed to have promoted it in the first place. In contrast, context-driven NHST requires researchers to specify the values they use within the NHST framework and to be able to justify these values based on the context of their research. Within CD-NHST applications, it is important to clarify what context means in specific and various research settings. Context can include (but is certainly not limited to) the nature of the research area (both major field and subfields), the research question, the sampling methodology, the study design, the sample size, the measurement of the data, the ethical implications regarding the research and sample, and the quality of the data, along with researcher judgment and experience.

The hypothesized effect size, due to theory or past research, is also fundamental in driving CD-NHST, as it can influence what alpha value is selected, the sample size needed, and the value of the null. Likewise, the desired level of power for a study is essential in both contributing context and requiring context. Specifically, desired or hypothesized effect size and desired level of power are fundamental in determining an appropriate alpha value (balance of errors) and null value. In general, effect sizes (e.g., to determine magnitude of effects) and power considerations (e.g., study design of detecting real effects) along with confidence intervals (e.g., to illustrate uncertainty around estimates) have long been championed as essential components and/or supplements to NHST (e.g., Denis, 2003; Harlow, 2010; Robinson & Levin, 2010). These and other broad contextual considerations are suggested in a matrix in Table 1. This matrix is presented only to stimulate additional and deeper research context considerations and how they relate to the alpha value, null value and range, and should not be viewed as an exhaustive list, or even worse as a replacement ritual.

In general, contextual aspects of research help guide researchers in deciding which statistical tools to use (e.g., CD-NHST, modeling, Bayesian, etc.) and how to implement these tools to evaluate research questions (see Gigerenzer, 2004). Indeed, as Abelson (1997) asserted:

DOES ONE SIZE FIT ALL? A CASE FOR CONTEXT-DRIVEN NHST

Table 1. Research context matrix

	Example Considerations	Critical Value α	Value of Null H_0	Range of Null $\Delta_{+/-}$
Research Question	i. Specific Hypothesis ii. No Hypothesis			
Study Type	i. Pilot ii. Exploratory iii. Confirmatory			
Measurement	i. Precise data (small variability/Reliable) ii. Noisy data (Large variability/ less reliable)			
Field of Research	i. Biological Psychology (e.g., precise biomarkers) ii. Clinical Psychology (e.g., self-report)			
Design	i. Experimental ii. Observational iii. Correlational			
Sampling	i. Probability-sampling ii. Non-probability sampling iii. Clinical sample			
Sample Size & Power	i. Small sample (related: underpowered) ii. Large sample (related: overpowered) iii. Level of power desired			
Cost	i. Type I error more costly relative to Type II error ii. Type II error more costly relative to Type I error iii. Cost of sample			
Seriousness	i. Type I error more serious relative to Type II error ii. Type II error more serious relative to Type I error			
Replication	i. Study is a replication of another study ii. Study is first of its kind iii. Study will probably not be replicated			
Previous Data	i. Previous results indicate for this study... ii. No previous results for this study...			
Effect size	i. Hypothesized magnitude of the effect, based on theory or past research.			

Note that each category is not mutually exclusive. For example, measurement variability is often closely related to field of research

Good methodologists should be open to the possibility that a method does not apply in a particular case, or that more information is required. Statistical methods are better conceived as options than as commandments. Each method has areas of application in which it is typically useful, and areas in which it is weak or open to criticism. (p. 14)

Earlier, Neyman remarked:

It may be useful to point out that although we are frequently witnessing controversies in which authors try to defend one or another system of the theory of probability as the only legitimate [one], I am of the opinion that several such theories may be and actually are legitimate, in spite of their occasionally contradicting one another. Each of these theories is based on some system of postulates, and so long as the postulates forming one particular system do not contradict each other and are sufficient to construct a theory, this is as legitimate as any other (Reid, 1982, p. 136).

Once the appropriate type of analysis is selected, researchers can use the context of the research to then guide and inform which values to use in the selected analysis. Although this holds for modeling and especially Bayesian analysis, which takes into account prior information, only the conventional NHST situation has been considered for the purposes of this paper. However, the field would benefit greatly from future work examining the issues regarding research context and other quantitative approaches such as statistical modeling (e.g., Harlow, 2010; McGrath, 2011; Rodgers, 2010a; 2010b).

Some may hold that NHST should be abandoned as an evaluative framework in science because it is often employed in a formulaic way. However, the argument presented here reveals that, regardless of how NHST may be commonly applied, it need not be used in a mechanistic way. Indeed, judging from original sources, it is questionable if null hypothesis testing or significance testing were ever designed to be used in the way they are applied today. Given the ability to designate alpha values and null values with context in mind, it is difficult to see why NHST is credited with being a mechanistic epistemological framework in the first place (see Rodgers, 2010a).

In closing, there are a number of errors that researchers must keep in mind when engaged in research. For instance, errors of the first kind are achieved when we incorrectly reject the null hypothesis whereas errors of the second kind are

achieved when we incorrectly accept the null hypothesis (Neyman & Pearson, 1933b). These are the familiar errors that must be considered when selecting alpha. Mosteller (1948, p. 61) proposed an error of a third kind, whereby we correctly reject the null, but for the wrong reason. Later Marascuilo and Levin (1970, p. 398) proposed that errors of the fourth kind are achieved when we correctly reject the null hypothesis but give the wrong interpretation. It is proposed here that errors of the infinite kind are achieved when we correctly or incorrectly reject or accept the null hypothesis, but do so without context. That is, a limitless supply of error is available when we conclude without context.

Acknowledgements

This research was supported in part by G20RR030883 from the National Institutes of Health.

References

- Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science*, 8(1), 12-15. doi: [10.1111/j.1467-9280.1997.tb00536.x](https://doi.org/10.1111/j.1467-9280.1997.tb00536.x)
- Aguinis, H., Werner, S., Abbott, J. L., Angert, C., Park, J. H. & Kohlhausen, D. (2010). Customer-centric science: Reporting significant research results with rigor, relevance, and practical impact in mind. *Organizational Research Methods*, 13(3), 515-539. doi: [10.1177/1094428109333339](https://doi.org/10.1177/1094428109333339)
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66(6), 423-437. doi: [10.1037/h0020412](https://doi.org/10.1037/h0020412)
- Baker, L. F. & Mudge, J. F. (2012). Making statistical significance more significant. *Significance*, 9(3), 29-30. doi: [10.1111/j.1740-9713.2012.00574.x](https://doi.org/10.1111/j.1740-9713.2012.00574.x)
- Beale, D. K. (1972). What's so significant about .05? *American Psychologist*, 27(11), 1079-1080. doi: [10.1037/h0038057](https://doi.org/10.1037/h0038057)
- Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology*, 36(3), 517-526. doi: [10.1111/j.1744-6570.1983.tb02233.x](https://doi.org/10.1111/j.1744-6570.1983.tb02233.x)
- Cochran, W. G. (1976). Early development of techniques in comparative experimentation. In D. B. Owen (Ed.), *On the history of statistics and probability* (pp. 1-26) New York, NY: Dekker.

- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, 49(12), 997-1003. doi: [10.1037/0003-066X.49.12.997](https://doi.org/10.1037/0003-066X.49.12.997)
- Cortina, J., & Landis, R. (2010). The Earth is *not* round ($p = .00$). *Organizational Research Methods*, 14(2), 332-349. doi: [10.1177/1094428110391542](https://doi.org/10.1177/1094428110391542)
- Cowles, M. & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist*, 37(5), 553-558. doi: [10.1037/0003-066X.37.5.553](https://doi.org/10.1037/0003-066X.37.5.553)
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- Denis, D. (2003). Alternatives to null hypothesis significance testing. *Theory & Science*, 4(1). Retrieved from http://theoryandscience.icaap.org/content/vol4.1/02_denis.html
- Feynman, R. (1985). "Surely you're joking, Mr. Feynman!": *Adventures of a curious character*. New York, NY: W. W. Norton & Company.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, 33, 503-513.
- Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed.). Edinburgh: Oliver & Boyd.
- Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver and Boyd.
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.) London, UK: Collins Macmillan.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587-606. doi: [10.1016/j.socec.2004.09.033](https://doi.org/10.1016/j.socec.2004.09.033)
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.
- Harlow, L. L., Mulaik, S., & Steiger, J. (Eds.) (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Harlow, L. L. (2010). On scientific research: The role of statistical modeling and hypothesis testing. *Journal of Modern Applied Statistical Methods*, 9(2), 348-358. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol9/iss2/4/>
- Hodges, J. L. & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society. Series B*

(*Methodological*), 16(2), 261-268. Available from
<http://www.jstor.org/stable/2984052>

Hubbard, R., & Ryan, P. A. (2000). The historical growth of statistical significance testing in psychology—and its future prospects. *Educational and Psychological Measurement*, 60(5), 661-681. doi: 10.1177/00131640021970808

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford.

Knapp, T. R. & Sawilowsky, S. S. (2001). Constructive criticisms of methodological and editorial practices. *The Journal of Experimental Education*, 70(1), 65-79. doi: 10.1080/00220970109599498

Marascuilo, L. A., & Levin, J. R. (1970). Appropriate post hoc comparisons for interaction and nested hypotheses in analysis of variance designs: The elimination of Type IV errors. *American Educational Research Journal*, 7(3), 397-421. Available from <http://www.jstor.org/stable/1161635>

McGrath, R. E. (2011). *Quantitative models in psychology*. Washington, DC: American Psychological Association.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2), 108-141. doi: 10.1207/s15327965pli0102_1

Morrison, D. E., & Henkel, R. E. (1969). Significance tests reconsidered. *American Sociologist*, 4(2), 131-140. Available from
<http://www.jstor.org/stable/27701482>

Mosteller, F. (1948). A k -sample slippage test for an extreme population. *The Annals of Mathematical Statistics*, 19(1), 58-65. Available from
<http://www.jstor.org/stable/2236056>

Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS ONE*, 7(2). doi: 10.1371/journal.pone.0032734

Murphy, K. R., & Myers, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum effect tests in the general linear model. *Journal of Applied Psychology*, 84(2), 234-248. doi: 10.1037/0021-9010.84.2.234

Murphy, K. R., & Myers, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, NJ: Erlbaum.

Neyman, J. (1942). Basic ideas and theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, 105(4), 292-327. doi: 10.2307/2980436

- Neyman, J. (1950). *First course in probability and statistics*. New York, NY: Henry Holt.
- Neyman, J. (1957). Inductive behavior as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique*, 25(1/3), 7-22. doi: 10.2307/1401671
- Neyman, J., & Pearson, E. S. (1933a). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, 231A, 289-337.
- Neyman, J., & Pearson, E. S. (1933b). The testing of statistical hypotheses in relation to probabilities a priori. *Proceedings of the Cambridge Philosophical Society*, 29(04), 492-510. doi: 10.1017/S030500410001152X
- Pearson, E. S. (1955). Statistical concepts in their relation to reality. *Journal of the Royal Statistical Society. Series B (Methodological)*, 17(2), 204-207. Available from <http://www.jstor.org/stable/2983954>
- Reid, C. (1982). *Neyman – from life*. Berlin, Germany: Springer.
- Robinson, D. H., & Levin, J. R. (2010). The not-so-quiet revolution: cautionary comments on the rejection of hypothesis testing in favor of a “causal” modeling alternative. *Journal of Modern Applied Statistical Methods*, 9(2), 332-339. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol9/iss2/2/>
- Rodgers, J. L. (2010a). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist*, 65(1), 1-12. doi: 10.1037/a0018326
- Rodgers, J. L. (2010b). Statistical and mathematical modeling versus NHST? There's no competition! *Journal of Modern Applied Statistical Methods*, 9(2), 340-347. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol9/iss2/3/>
- Sawilowsky, S. S. (2003). Deconstructing arguments from the case against hypothesis testing. *Journal of Modern Applied Statistical Methods*, 2(2), 467-474. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol2/iss2/19/>
- Selvin, H. C. (1958). Reply to Gold's comment on “A critique of tests of significance”. *American Sociological Review*, 23(1), 86. Available from <http://www.jstor.org/stable/2089071>
- Serlin, R. C. (1987). Hypothesis testing, theory building, and the philosophy of science. *Journal of Counseling Psychology*, 34(4), 365-371. doi: 10.1037/0022-0167.34.4.365

DOES ONE SIZE FIT ALL? A CASE FOR CONTEXT-DRIVEN NHST

Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good-enough principle. *American Psychologist*, 40(1), 73-83. doi: [10.1037/0003-066X.40.1.73](https://doi.org/10.1037/0003-066X.40.1.73)

Skipper, J. K., Guenther, A. L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *The American Sociologist*, 2(1), 16-18. Available from <http://www.jstor.org/stable/27701229>

Yule, G. U., & Kendall, M. G. (1950). *An introduction to the theory of statistics* (14th ed.). London, UK: Griffin.

A Two and More Independent Factor Form of the Wilcoxon-Mann-Witney Test, Extendable to Other Permutation-Based Tests

Christopher H. Holland

King's College London
London, United Kingdom

Tom Holland

Triveritas Ltd.
Brampton, Cumbria, United Kingdom

The Wilcoxon-Mann-Witney test is extended to account for a second independent factor. The new test statistic's probability mass function and normal approximation are derived. Critical-values for balanced, unbalanced, and large sample designs are given. The immediate extension of this method to a wide range of non-parametric tests is explained.

Keywords: Two factor nonparametric test, Wilcoxon-Mann-Witney test, nuisance factors, rank, order, permutation

---

Introduction

The Wilcoxon-Mann-Witney test (WMWt) is a widely used technique for data analysis in which a natural ordering is possible. For example, it may be possible to order or rank the subjective degree of inflammation or pain; yet objective measurement of the inflammation or pain may be impossible. The WMWt is simple, robust and powerful; it has a minimum asymptotic relative efficiency (ARE or Pitman efficiency) of 95% compared to Students' t test, the most efficient test possible under ideal conditions (Conover, 1999). With small samples, heteroscedastic, and non-normal data the WMWt can have much greater power than Students t test (Blair, Higgins, & Smitely, 1980). The WMWt is widely taught and used because of these properties – it also has great intuitive appeal to experimenters' common sense.

However, it is common that experiments are run in which an extraneous factor (such as sex of the subject or strain of animal) is included which is not of direct experimental interest and may confound the result. When this occurs, it is unreasonable to ignore the 'nuisance' factor in the analysis of ranked/ordered data

---

Mr. Christopher Holland is a graduate medical student in the clinical portion of his training. Email him at: christopher.c.holland@kcl.ac.uk.

on the effect of the treatment. For example, in assessing the effect of a treatment on pain or inflammation by ranking of pain or inflammation in both sexes, it would be unreasonable to simply ignore the sex of the subjects. This is because the sex of an experimental subject substantially influences both the nature and degree of an inflammatory response and the perception of pain in man and animals (reviewed by Berkley, Zalcman, & Simon, 2006), and can thus confound any result.

A practical method of extending the WMWt is proposed so that a nuisance factor which divides the sample into two distinct sets for which direct comparison of subjects between sets is precluded (such as subject sex) can be accounted for in the detection of a significant effect of the explanatory variable upon the dependent variable. As such, this nuisance factor can be excluded from influencing the conclusions drawn. The proposed extension unifies an overall analysis of otherwise subsetted data and so offers increased statistical power and avoids conflicting conclusions between data subsets. It also avoids the ambiguous situation where a series of tests on small subsets of the data by the simple WMWt may give different results (e.g. the males show a significant treatment effect which is not shown by the females). Measuring interactions between factors (i.e. detecting if there is a difference in the degree of response of the males compared to the females) is not possible in this simple formulation given here. However, a substantially more complex extension of our techniques makes this possible (theoretical approach outlined in Holland, 2011).

Heuristic Development of the Statistical Model

Consider the one-tailed Mann-Witney formulation of the U statistic (Sprent & Smeeton, 2001). This test is used to compare two groups of subjects, made distinct by some factor (such as in an animal study, a treated group and a control group) for which the subjects in the two groups can be ordered in some feature (for example an experimental parameter). For the remainder of the article we refer to the two groups as a control group (denoted c) and a treated group (t) for convenience. The purpose is to deduce whether there is a statistically significant difference in this feature between the two groups. To obtain a U statistic from the data, one takes each element of control group and counts the number of samples in the treated group (t) that show less of the feature. The final U value is obtained by summing this count for each element of the control group.

Hence	<i>least</i>	tctc	<i>most</i>	$U = 4$
		cttc		$U = 4$
while		tctc		$U = 5$

To assess statistical significance, we wish to compare this value with that expected assuming the null hypothesis. Under the null hypothesis that there is no difference between the two groups, every ordering of samples is equally likely. Thus the key step in assigning significance to a particular U -value obtained from an experiment is to identify the probability mass function under the null hypothesis by quantifying the number of different permutations of the given numbers of treated and controls that give rise to each possible value of U . A result is significant if the U -value falls beyond a certain point in the extreme tail of this probability mass function.

The usual assumption underpinning the WMWt applies to this work, namely: the samples are independent of each other. So this is a test simply for a difference in location or, equivalently, a difference in the mean (if it exists) and median between the two groups.

The proposed test procedure accounts for the scenario in which the subjects in the sample are additionally divided by a second factor into two sets (e.g. male and female) where comparison of individuals between these two sets is precluded, effectively making it impossible to calculate the U -statistic of the whole sample. Here, the first two U statistics are calculated by first considering the two sets separately and calculating the U statistic of each with respect to the feature of interest (labeled U_1 and U_2). For example, if sex is this second factor, a U_1 statistic for the males is produced and a U_2 statistic for the females is produced in the usual way, considering the two sets as completely separate; as such, no comparison of males to females or females to males is made.

However, if the analysis ends here, then two different statistics are produced despite both being the result of an identical experimental design, and the data has effectively been subsetting. This causes a number of difficulties. For instance, due to the reduction in the group sizes from the original, each would have reduced statistical power for a given overall sample size. There is also the potential for the two results to have similar but conflicting implications (for example just managing to achieve significance in one set, while just failing to achieve significance in the other), making it unclear as to the overall conclusion.

To avoid such problems, it is proposed that instead of ending the analysis there, the two separately obtained U -values (U_1 and U_2) are added together to form a *combined U -value* U_C . This approach has the very appealing property that each comparison of any individual pair contributes exactly the same amount of information to the overall test statistic as any other comparison, and so any difference of sample size requires no correction or weighting of the contributing U_1 or U_2 values to U_C .

To analyze this result further, the U_C is then compared to the distribution of U_C statistics expected under the null hypothesis to infer the statistical significance of any overall difference between the two groups, similarly to the one-factor case.

Establishing the number of distinct permutations giving rise to each possible U_C value, and thus establishing the probability mass function of the U_C statistic under the null hypothesis, is the real substance of this article. We then use this to tabulate critical values for all balanced two-factor study designs up to a group size of 10 for one- and two-tailed designs. For unbalanced and larger group sizes we give a computer algorithm that gives an exact probability mass function and a simple normal approximation method.

Methodology

Revisiting the Analysis for a Single Factor U Statistic

First, to aid in understanding the two set case, we reexamine the well-established case of a single set of directly comparable subjects divided into two groups: a treated group with m subjects and a control group with n subjects. Consider the scenario in which, for an appropriate feature, the ordering of the members of the two groups is found and the U -value is calculated. The aim is to establish whether this U -value represents a statistically significant difference between the two groups by comparing it to what is expected under the null hypothesis: that there is no overall difference in the feature between the two groups.

Under the null hypothesis, this definition implies that, for any given subject, its position in the ordering is not influenced at all by which group it is from. Hence, if the null hypothesis is true, each of the orderings of the subjects is equally likely. Furthermore, we may ignore the internal orderings of both the treated and control groups and so just consider the distinct permutations of m indistinguishable t's and n indistinguishable c's as each such permutation is equally likely. For example, in case $m = 2, n = 3$, there would be an equal probability of an experiment giving rise to the ordering 'tctcc' as 'ctctc' (or any other distinct permutation of two t's and three c's, for that matter). This is because such a permutation represents $m!n!$ different subject orderings. As this number is constant for all permutations and each ordering is equally likely, each of the permutations is each equally likely under the null hypothesis.

Each of these distinct permutations has an associated U -value. The next step is therefore to establish the number of distinct permutations that give rise to each U -value. The function $f_1(m, n | r)$ is introduced for this purpose. This represents, for

treated and control groups of size m and n respectively, the number of permutations that give a U -value of r . For example, in an experiment involving 2 treated and 3 controls,

$$f_1(2, 3 | r) = 1, 1, 2, 2, 2, 1, 1$$

for $r = 0, 1, 2, 3, 4, 5, 6$, respectively. So, for instance, by looking at the 5th element in the sequence, we can deduce that there are two permutations giving rise to a U -value of 4. These are tcctc and ctcc.

As previously established under the null hypothesis, the distinct permutations are each equally likely; this shows that, in such conditions, it would be twice as likely to obtain a U -value of 3 as a U -value of 1 (for example). As such, for fixed m and n , the value of $f_1(m, n | r)$ represents the relative probability of achieving the different values of U (represented by r) under the null hypothesis. Therefore, if this function is normalized for fixed m and n , we obtain the probability mass function for U under the null hypothesis:

$$p_{1:m,n}(r) = \frac{f_1(m, n | r)}{{}^{m+n}C_m} . \quad (1)$$

This uses the property

$$\sum_{i=1}^{mn} f_1(m, n | i) = {}^{m+n}C_m . \quad (2)$$

Hence as it effectively yields the distribution of U -values expected under the null hypothesis, $f_1(m, n | r)$ is the key function in establishing whether an experimentally obtained U -value represents a statistically significant difference between the two groups or not. No closed form of $f_1(m, n | r)$ exists, but it satisfies recursion relations which allow its calculation using a computer (some such recursive properties are reviewed in Di Bucchianico (1999)).

Due to the importance of $f_1(m, n | r)$, we discuss some of its properties. First, we note that there are no permutations yielding a U -value which is negative or in excess of mn . As such, $f_1(m, n | r) = 0$ unless $0 \leq r \leq mn$. Secondly, the sum over all possible values of $f_1(m, n | r)$ is given by (2).

Finally, for fixed m, n , $f_1(m, n | r)$ is symmetric in r about the central value, $mn/2$, so that

TWO-FACTOR WILCOXON-MANN-WHITNEY

$$f_1(m, n | r) = f_1(m, n | mn - r) . \quad (3)$$

This also means that both the mean and median value of U under the null hypothesis are $mn/2$.

In order to find $f_1(m, n | r)$ for given m, n , use the method suggested by Theorem 2.6 in Di Bucchianico (1999), which we derived without the use of the correspondence with the restricted partition function. This utilizes the relations

$$\begin{aligned} f_1(m, n | r) &= 0 \quad \text{for } r < 0, r > mn, m < 0, n < 0 \\ f_1(m, n | 0) &= 1 \quad \text{for } m \geq 0, n \geq 0 \\ f_1(m, n | r) &= f_1(m-1, n | r-n) + f_1(m, n-1 | r) \quad \text{for } 1 \leq r \leq mn, m \geq 1, n \geq 1 \end{aligned}$$

Using these formulae, it is possible to specify $f_1(m, n | r)$ for any finite values of m, n , or r . The Matlab 7.8.0 program which we used to find $f_1(m, n | r)$ for given m and n and for the possible values of r is given in Appendix 1. This was able to calculate $f_1(20, 20 | r)$ for all possible r practically instantaneously, and $f_1(100, 100 | r)$ within 100 seconds using a home laptop computer.

Having obtained the probability mass function under the null hypothesis, it is a simple matter to analyze the statistical significance of a given result. For example, it can be used to calculate the P -value once an experimental U -value is obtained or a specified confidence interval for the null value.

However, as the values of m and n become very large, it becomes unfeasibly onerous even for a computer to calculate $f_1(m, n | r)$ exactly. In that case, approximations may be of greater practical use. For sets where each group size exceeds 20, the normal approximation is deemed sufficiently accurate for most usual cases. This involves approximating the probability mass function $p_{1;m,n}(r)$ by a normal distribution, with mean $mn/2$ and standard deviation

$$\sigma_U = \sqrt{\frac{mn(m+n+1)}{12}} . \quad (4)$$

Defining the Frequency Distribution of the U_c Statistic

As outlined above, the aim is to extend the analysis to the case where there are two sets which differ in some factor other than the experimental parameter, such that an element of one set cannot be reasonably compared to an element of the other set. A *combined U-value* U_c is formed by finding a U -statistic for each set separately and

then adding the two values. As in the single set case, the goal is to determine the statistical significance of an experimentally-obtained U_C -value.

By identical reasoning as in the single set case, if the null hypothesis holds, then each combination of distinct permutations of the treated and control subjects in each set is equally likely. As before, this observation leads to the construction of a function that enumerates the number of permutation combinations that give rise to a given U_C -value for specified group sizes.

Start by establishing a function: $F_2(m, n; p, q | r; s)$. This function gives the number of permutation combinations which simultaneously achieve: a U_1 -value of r in set one (with treated group size m and control group size n) and a U_2 -value of s in set two (with treated group size p and control group size q). Due to the independence of the two sets, this is the product of the number of permutations in set 1 giving a U -value of r and the number of permutations in set 2 giving a U -value of s . Alternatively, using the notation of the previous section:

$$F_2(m, n; p, q | r, s) = f_1(m, n | r) * f_1(p, q | s) . \quad (5)$$

This can be represented, for given m, n, p , and q , as a matrix with the row denoted by r and the column by s . An example of this is given in Table 1, where it is detailed for $F_2(2, 2; 2, 3 | r; s)$.

From $F_2(m, n; p, q | r; s)$, the function can be created giving the number of permutation combinations giving rise to a U_C -value of k : $f_2(m, n; p, q | k)$. This is the equivalent of $f_1(m, n | r)$ for the two set case. It is constructed by adding together all elements of $F_2(m, n; p, q | r; s)$ with the specified values of m, n, p, q where the sum of r and s is equal to k , as can be seen in (6) below:

Table 1. The two-dimensional U frequency array for an $F_2(2, 2; 2, 3 | r, s)$ study design, illustrating how the sum of each trailing diagonal gives the total number of ways of achieving each value of U_C ($U_C = 3$ cells are identified explicitly)

U_1 value (r)	U_2 value (s) Individual frequency	0	1	2	3	4	5	6
		1	1	2	2	2	1	1
0	1	1	1	2	2*	2	1	1
1	1	1	1	2*	2	2	1	1
2	2	2	2*	4	4	4	2	2
3	1	1*	1	2	2	2	1	1
4	1	1	1	2	2	2	1	1

Note: * $U_C = 3$

TWO-FACTOR WILCOXON-MANN-WHITNEY

$$f_2(m, n; p, q | k) = \sum_{i+j=k} F_2(m, n; p, q | i; j) . \quad (6)$$

In the matrix representation, this is simply adding the entries along the appropriate trailing diagonal, which is shown in Table 1 for $F_2(2, 2; 2, 3 | r; s)$ for the U_C value of $k = 3$. By summing the indicated numbers, we find that $f_2(2, 2; 2, 3 | 3) = 7$. To illustrate the point, these seven permutation combinations for males and females respectively may be identified as:

tctc and ccctt (U -values 3 & 0 respectively); tcct and cctct (2 & 1); ctte and cctct (2 & 1); ctct and ccttc (1 & 2); ctct and ctctc (1 & 2); cctt and ctctc (0 & 3); cctt and tcct (0&3)

By carrying out similar sums along the other trailing diagonals,

$$f_2(2, 2; 2, 3 | k) = 1, 2, 5, 7, 10, 10, 10, 7, 5, 2, 1$$

for $k = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$, respectively.

Several properties of $f_2(m, n; p, q | k)$ follow directly from the properties of $f_1(m, n | r)$. For instance, the function is non-zero only for $0 \leq k \leq (mn + pq)$, and the overall number of permutations regardless of U_C -value (i.e. the sum of $f_2(m, n; p, q | k)$ over all k for given group sizes) is

$$\sum_{i=0}^{mn+pq} f_2(m, n; p, q | i) = {}^{m+n}C_m * {}^{p+q}C_p . \quad (7)$$

Somewhat less trivially, $f_2(m, n; p, q | k)$ is also symmetric in k around the central value $(mn + pq)/2$. This is seen as follows:

$$\begin{aligned} f_2(m, n; p, q | (mn + pq) - k) &= \sum_{i+j=(mn+pq)-k} f_1(m, n | i) f_1(p, q | j) \\ &= \sum_{(mn-i)+(pq-j)=k} f_1(m, n | mn-i) f_1(p, q | pq-j) \\ &= \sum_{i'+j'=k} f_1(m, n | i') f_1(p, q | j') \\ &= f_2(m, n; p, q | k) \end{aligned}$$

On line 3, dummy variables i and j were replaced by dummy variables $i' = pq - i$ and $j' = pq - j$, respectively. The mean and median of $f_2(m, n; p, q | k)$ are therefore both $(mn + pq)/2$. Finally, as for $f_1(m, n | r)$, we can use $f_2(m, n; p, q | k)$ to construct the probability mass function for U_C :

$$p_{2;m,n;p,q}(k) = \frac{f_2(m, n; p, q | k)}{\binom{m+n}{m} \binom{p+q}{p}}. \quad (8)$$

These properties allow a U_C -value to be analyzed using the same approach as the U -value in the single set case, just using $f_2(m, n; p, q | k)$ instead of $f_1(m, n | r)$ to, for example, calculate the P -value of a given U_C or a confidence interval under the null hypothesis. This is discussed further in the section below. The recursive Matlab program we used to find $f_2(m, n; p, q | k)$ for given m, n, p , and q , and for the possible values of k is given in Appendix 2.

Again, it is possible to use a normal approximation to $p_{2;m,n;p,q}(k)$ for large group sizes, ideally all groups in excess of 10. This is accomplished simply by the distribution resulting from the addition of normally distributed random variables of the two separate normal approximations of the two sets. This results in a normal distribution with mean $(mn + pq)/2$ and standard deviation

$$\sigma_{U_C} = \sqrt{\frac{mn(m+n+1) + pq(p+q+1)}{12}}. \quad (9)$$

Table 2. The values that U_C must Equal or Exceed to achieve significance at the given confidence level for one-tailed and two-tailed tests

Group Size	One-Tailed		Two-Tailed	
	95%	99%	95%	99%
1	-	-	-	-
2	8	-	-	-
3	15	17	2, 16	0, 18
4	25	28	5, 27	3, 29
5	37	41	11, 39	7, 43
6	52	57	18, 54	13, 59
7	68	75	26, 72	20, 78
8	87	96	37, 91	29, 99
9	108	119	49, 113	39, 123
10	132	144	62, 138	52, 148

Results

Using the Exact Distribution of U_C

As discussed above, obtaining the probability mass function under the null hypothesis allows us to establish confidence intervals for the null hypothesis. This gives us a range of U_C -values for which the null hypothesis can be rejected for a specified statistical significance (α). This gives us critical values of U_C , which are the highest (or lowest) value of U_C such that the null hypothesis cannot be rejected. The most commonly used values of α are 0.05 and 0.01, which give a 95% and 99% confidence interval, respectively, for either one- or two-tailed tests. As such, the critical values of U_C for all experiments with uniform group sizes (i.e. $m = n = p = q$) up to 10 are given in Table 2 for these confidence intervals. A dash is used for circumstances in which the test can never yield a statistically significant result.

To calculate the critical values for other cases, such as unbalanced designs or for data in which group sizes are larger than 10, a program such as that in Appendix 2 can be used to find the $f_2(m, n, p, q | k)$ function from which the probability mass function can be obtained. Any confidence interval under the null hypothesis can then be constructed and, similarly, an exact P -value can be calculated.

Using the Asymptotic Distribution of U_C

Where m, n, p , and q are greater than 10, the normal approximation can be used (derived from Campbell (1974), including continuity correction). The distribution of z follows the Normal distribution and the one- and two-tailed critical values can be taken directly from standard tables:

$$z = \frac{\left(U_C \pm \frac{1}{2} - \frac{(mn + pq)}{2} \right)}{\sqrt{\frac{1}{12} (mn(m+n+1) + (pq(p+q+1)))}} \quad (10)$$

The values of m, n, p , and q at which this approximation becomes workably accurate for the 95% confidence level, both one- and two-tailed, are investigated using tables that give 3 significant figures for z . The exact p value is computed for all combinations of group sizes up to 10 (i.e. 4 groups each of 10 or less samples) for all U_C values possible, and the U_C value which just exceeds the 95% confidence limit is found. Hence the next U_C value is obtained closer to the mean than this, the

highest non-significant U_C value. Then, compare this lowest significant U_C highest non-significant U_C pair to the value given by the normal approximation form of the test. The normal approximation test is very close to the exact value unless one or more groups have a group size of one. Except for this extreme condition, the one-tailed test normal approximation never gives a significant result as non-significant. However, in 31 of 1035 one-tailed cases, the normal approximation gives as significant a U_C that the exact test gives as just non-significant. The group sizes that give a false positive are given in Table 3; all are the result of rounding errors, and the normal approximation gives complete concordance with the exact test working to 7 figure accuracy. In the two-tailed test, none of the highest non-significant U_C values give a significant finding, but the approximate methods fails to detect significance in 91 of 1035 two-tailed test. The group sizes that give a false negative are given in Table 4.

We think this remarkable accuracy of the normal approximation is the result of the action of the Central Limit Theorem when combining two mass functions that are themselves fairly close to normal. Practically, it means that so long as $m+nC_m * p+qC_p > 20$ (one-tailed) and 41 (two-tailed), then (given that the 95% confidence limit is an arbitrary cut-off point) the normal approximation is a practical method for all study designs that do not use group sizes of 1.

Table 3. One-tailed 95% confidence limit; these are the groups sizes that are just not significant at the exact 95% confidence limit, but in the normal approximations are just significant ($z = \pm 1.65$ to 3 significant figures for all them)

2,	4:	8,	9	3,	4:	7,	10
2,	5:	4,	9	3,	5:	7,	9
2,	4:	9,	10	4,	10:	7,	10
2,	4:	9,	9	4,	7:	6,	10
2,	5:	5,	7	4,	5:	6,	9
2,	10:	5,	5	4,	6:	5,	10
2,	10:	7,	7	4,	9:	5,	8
2,	7:	10,	10	4,	5:	4,	5
2,	2:	6,	6	4,	4:	10,	10
2,	4:	4,	6	4,	8:	9,	9
2,	7:	3,	6	5,	5:	6,	8
2,	8:	4,	4	5,	8:	10,	10
2,	7:	2,	10	6,	8:	8,	10
2,	3:	7,	8	6,	9:	7,	7
3,	9:	5,	7	9,	10:	9,	10
3,	8:	4,	5	-	-	-	-

Note: The order in each pair is immaterial, and the order of the pairs themselves is also immaterial (so 9,8:2,4; 8,9:2,4; 9,8:4,2; 8,9:4,2; 4,2:8,9; 4,2:9,8, and 2,4:9,8 are all equivalent to the first entry in the table, 2,4:8,9)

TWO-FACTOR WILCOXON-MANN-WHITNEY

Table 4. Two-tailed 95% confidence limit; these are the group sizes that are just significant at the exact 95% confidence limit, but in the normal approximations are just not significant ($z < \pm 1.65$ to 3 sf., but for all $z > \pm 1.520$)

2,	2:	2,	5	2,	6:	3,	6	3,	3:	6,	6	4,	6:	4,	10
2,	2:	2,	9	2,	6:	3,	9	3,	3:	9,	9	4,	7:	6,	9
2,	2:	3,	3	2,	6:	5,	5	3,	4:	9,	10	4,	7:	7,	10
2,	2:	3,	5	2,	6:	5,	8	3,	5:	3,	6	4,	8:	9,	10
2,	2:	4,	6	2,	6:	6,	6	3,	5:	7,	10	4,	9:	8,	8
2,	2:	5,	8	2,	6:	6,	9	3,	6:	3,	9	4,	10:	4,	10
2,	3:	2,	10	2,	6:	8,	8	3,	6:	5,	5	4,	10:	6,	8
2,	3:	3,	9	2,	6:	10,	10	3,	6:	6,	10	5,	5:	6,	9
2,	3:	4,	4	2,	7:	3,	3	3,	6:	7,	7	5,	6:	6,	8
2,	3:	5,	9	2,	7:	4,	4	3,	6:	7,	8	5,	8:	7,	8
2,	3:	6,	8	2,	7:	6,	8	3,	7:	7,	7	5,	9:	6,	7
2,	3:	8,	9	2,	8:	2,	8	3,	7:	8,	9	5,	9:	8,	9
2,	4:	2,	8	2,	8:	6,	10	3,	8:	3,	8	5,	9:	9,	9
2,	4:	5,	6	2,	8:	7,	7	3,	8:	6,	6	5,	10:	8,	8
2,	4:	5,	10	2,	9:	4,	5	3,	8:	7,	7	6,	7:	6,	9
2,	4:	6,	10	2,	9:	5,	6	3,	8:	10,	10	6,	7:	9,	9
2,	5:	3,	10	2,	9:	6,	9	3,	9:	5,	6	6,	8:	6,	10
2,	5:	4,	7	2,	9:	8,	10	3,	9:	5,	8	6,	10:	10,	10
2,	5:	4,	8	2,	10:	3,	3	3,	9:	9,	9	7,	7:	10,	10
2,	5:	5,	7	2,	10:	3,	4	3,	10:	4,	7	7,	8:	7,	9
2,	5:	6,	7	2,	10:	4,	10	4,	4:	6,	8	7,	10:	8,	9
2,	5:	8,	10	2,	10:	5,	5	4,	5:	5,	8	8,	10:	9,	10
2,	5:	2,	8	3,	3:	5,	8	4,	5:	7,	10	-	-	-	-

Note: As described in Table 3, the order in each pair is immaterial, as is the order of the pairs themselves

Conclusion

The explanatory factorial approach taken here can be directly extended to any of the large family of tests in which the full extent of all possible test statistics is created and the most extreme tail of that distribution defined as the critical region. This includes all the WMWt family of tests (e.g. Jonckheere-Terpstra Test, Kruskal-Wallis Test), Kolmogorov-Smirnov-type tests (e.g. Conover test, Birnbaum-Hall Test), and Pitman permutation tests (Conover, 1999). The generic approach is:

1. Separate the experimental subjects across all factors and establish the test statistic for the two factors of experimental interest
2. Sum the test statistics across all the nuisance factors to get a combined test statistic

3. Derive the probability mass function for the combined statistic under the null hypothesis
4. Establish if the combined test statistic from step 2 is in the extreme tail of the distribution derived in step 3

This method also has direct extension to factors that hold 3 or more states – if, for example, one ran an experiment with three or more different strains of animal, in several age groups of clinical patients, or with clinical results from three or more different hospitals, it would then be possible to form a combined U -value by summing the three separate results. This could be compared with a similarly-derived probability mass function for this value under the null hypothesis.

There are weaknesses in the method. If one sex responds in a quantitatively different manner to the other (interaction of treatment with another factor), this is not measured. Subsetting the data on the nuisance factor and doing separate analyses on U_1 and U_2 might make one suspect interaction. Ties are not obviously incorporated in this simple formulation, although the method for correcting for ties in the derivation of U_1 and U_2 should be applicable to give a U_C that is unaffected by ties in U_1 and U_2 . We are currently developing this approach to ordinal contingency table data, in which very extensive ties are usual.

There are other approaches that might be adopted to achieve the same ends. Substituting normal scores for the ranks and then using a procedure such as ANOVA would be one. A Shirley test approach ([Williams, 1986](#)) of using the ranks themselves as though they were interval data and using a parametric procedure might be possible. When the test statistic for each subset of the data approximates to an established distribution (at it does in the Kruskal-Wallis test with the X^2 distribution), then combining those individual statistics may be possible for a test of the complete data set. Fisher's combined probability test and Weighted Z methods are inappropriate as they assume the probability for each table is uniformly distributed on the interval $[0, 1]$, but their p values can only hold discrete values. However, the approach we advocate has these advantages:

1. It is valid for very small groups; with 2 sexes, a 2 factor study with just two subjects for each sex and treatment can give significant results
2. It is robust with any distribution in the data; the groups do not even have to be drawn from the same distribution family, simply independently of each other

TWO-FACTOR WILCOXON-MANN-WHITNEY

3. Experimenters can clearly comprehend the rationale behind the test
4. The meaning of the result is clear
5. It is computationally simple

There are numerous practical situations in which experimenters may want to use a non-parametric test such as the WMWt, but there exists in their study design an unavoidable ‘nuisance factor’ which precludes the simple application of the test. Animal experiments, human clinical trials data, and psychological tests that include data from both sexes or can be age-stratified are all examples of large classes of such experiments. It is very rare that one can unequivocally exclude sex or age as a possible latent factor in such experiments, so it would be prudent to adopt as routine such methods as these. In the field in which one of us works – toxicology – it is common for experiments to be conducted with small groups sizes ($n = 3$ or 4) for studies involving primates or dogs (there are ethical objections to primate experiments with large group sizes). Doing the analyses on the sexes separately markedly limits the power of such experiments. However, doing tests including sex as a factor in the analysis of the complete data sets has a substantial beneficial effect on the power of such experiments. Simulation shows that with group sizes of 3 and using data from both sexes and all dose groups the power then approaches that of larger studies with group sizes of 10 in which the data for the sexes is not combined (Holland, 2011).

References

- Berkley K. J., Zalcman, S. S., & Simon, V. R. (2006). Sex and gender differences in pain and inflammation: A rapidly maturing field. *American Journal of Physiology – Regulatory, Integrative and Comparative Physiology*, 291(2), R241-R244. doi: [10.1152/ajpregu.00287.2006](https://doi.org/10.1152/ajpregu.00287.2006)
- Blair, R. C., Higgins, J. J., & Smitely, W. D. G. (1980). On the relative power of the U and t tests. *British Journal Mathematical and Statistical Psychology*, 33(1), 114-120. doi: [10.1111/j.2044-8317.1980.tb00783.x](https://doi.org/10.1111/j.2044-8317.1980.tb00783.x)
- Campbell, R. C. (1974). *Statistics for biologists* (2nd ed.). London: Cambridge University Press.
- Conover, W. J. (1999). *Practical nonparametric statistics* (3rd ed.). New York, NY: John Wiley and Sons.

Di Buccianico, A. (1999). Combinatorics, computer algebra and the Wilcoxon-Mann-Witney test. *Journal of Statistical Planning and Inference*, 79(2), 349-364. doi: [10.1016/S0378-3758\(98\)00261-4](https://doi.org/10.1016/S0378-3758(98)00261-4)

Holland, T. (2010). *A study of the methods used in toxicological pathology* (Fellowship thesis). Retrieved from Royal College of Veterinary Surgeons' Library, London (T/712).

Sprent, P., & Smeeton, N. C. (2001). *Applied non-parametric statistical methods* (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Williams, D. A. (1986). A note on Shirley's nonparametric test for comparing several dose levels to a non-zero control. *Biometrics*, 42(1), 183-186. doi: [10.2307/2531254](https://doi.org/10.2307/2531254)

Appendix 1

Matlab Program Calculating $f_1(m, n, r)$ for specified m, n

$f_1(m, n | r)$ is specified for given inputs m, n by $fvector[i] = f_1(m, n | i - 1)$.

```
function [ fvector ] = U1( ntreat, ncntl )

m=ntreat;
n=ncntl;

farray=zeros(m+1,n+1,m*n+1);

farray(1,1,1)=1;

fvector=farray(1,1,1);

for j=2:(m+1)
    farray(j,1,1)=1;
end;

for k=2:(n+1)
    farray(1,k,1)=1;
end;

for k=2:(n+1)
    for j=2:(m+1)
        for l=(1:(k-1)*(j-1)+1)
            sum=0;
            for h=1:min(l,k)
                sum=sum+farray(j-1,h,l-h+1);
            end;
            farray(j,k,l)=sum;
        end;
    end;
end;

fvector=zeros(m*n+1,1);

for l=1:(m*n)+1
    fvector(l,1)=farray(m+1,n+1,l);
end;

end
```

Appendix 2

Matlab Program Calculating $f_2(m, n; p, q | k)$

$f_2(m, n; p, q | r)$ is specified for given inputs m, n, p, q by $fvector[i] = f_2(m, n; p, q | i - 1)$.

```
function [ fvector ] = U2( ntreat1, ncnt11, ntreat2, ncnt12 )

m1=ntreat1;
n1=ncnt11;
m2=ntreat2;
n2=ncnt12;

m=max(m1,m2);
n=max(n1,n2);

farray=zeros(m+1,n+1,m*n+1);

farray(1,1,1)=1;

fvector=farray(1,1,1);

for j=2:(m+1)
    farray(j,1,1)=1;
end;

for k=2:(n+1)
    farray(1,k,1)=1;
end;

for k=2:(n+1)
    for j=2:(m+1)
        for l=(1:(k-1)*(j-1)+1)
            sum=0;
            for h=1:min(l,k)
                sum=sum+farray(j-1,h,l-h+1);
            end;
            farray(j,k,l)=sum;
        end;
    end;
end;

fvector1=zeros(m1*n1+1,1);
```

TWO-FACTOR WILCOXON-MANN-WHITNEY

```

for l=1:(m1*n1)+1
    fvector1(l,1)=farray(m1+1,n1+1,l);
end;

fvector2=zeros(m2*n2+1,1);

for l=1:(m2*n2)+1
    fvector2(l,1)=farray(m2+1,n2+1,l);
end;

fmatrix=zeros(m1*n1+1,m2*n2+1);

for j=1:m1*n1+1
    for k=1:m2*n2+1
        fmatrix(j,k)=fvector1(j,1)*fvector2(k,1);
    end;
end;

fvector=zeros(m1*n1+m2*n2+1,1);

for j=1:m1*n1+m2*n2+1
    sum=0;
    for k=max(1,j-m1*n1):m2*n2+1
        if k-j>0
            break;
        end;
        sum=sum+fmatrix(j-k+1,k);
    end;
    fvector(j,1)=sum;
end;

end

```

Parametric Tests for Two Population Means under Normal and Non-Normal Distributions

Diep T. Nguyen
University of South Florida
Tampa, FL

Eun Sook Kim
University of South Florida
Tampa, FL

Patricia Rodriguez de Gil
University of South Florida
Tampa, FL

Anh Kellermann
University of South Florida
Tampa, FL

Yi-Hsin Chen
University of South Florida
Tampa, FL

Jeffrey D. Kromrey
University of South Florida
Tampa, FL

Aarti Bellara
University of Connecticut
Storrs, CT

A simulation study was conducted to explore the performance of the independent means t -test, Satterthwaite's approximate t -test, and the conditional t -test under various conditions. Type I error control and statistical power of these testing approaches were examined and guidance provided on the proper selection among them.

Keywords: Type I error control, statistical power, parametric tests, independent means t -test, Satterthwaite's approximate t -test, conditional t -test

---

Introduction

In elementary statistics courses, the independent means t -test is followed by a discussion of statistical assumptions, robustness, Type I error control, and power. At the time of writing, the independent means t -test has been widely used in almost every discipline to this day. A search completed in June of 2014 with the key words "independent means t -test", with time period between 2013 and 2014, returned 1,740 articles from the Google Scholar database (excluding citation and patents) and 605 articles from the Web of Science database. Among the 605 articles in Web of Science, 170 out of the 202 most recent articles mentioned in the abstract that these studies utilized the independent means t -test.

Eun Sook Kim is an Assistant Professor in the Department of Educational and Psychological Studies. Email her at: ekim3@usf.edu.

PARAMETRIC TESTS FOR TWO POPULATION MEANS

The independent means t -test relies on the assumptions of population normality and equal variances (homoscedasticity). Alternative approaches such as the Satterthwaite's approximate t -test (Satterthwaite's test hereinafter) relax these assumptions, approximating the t distribution and the corresponding degrees of freedom. Although the independent means t -test is “the most powerful unbiased test” (Bridge & Sawilowsky, 1999, p. 229) for detecting true mean differences under the assumption of normality, statisticians to date are still evaluating the various conditions and factors for which this test is robust under the violation of the equality of variances assumption and departures from normality.

Controversy about the Independent Means t -Test

Many statistical textbooks (e.g., Cody & Smith, 1997; Schlotzhauer & Littell, 1997) continue recommending what Hayes and Cai (2007) call the “conditional decision rule” (p. 217), that researchers screen their samples for variance homogeneity by conducting preliminary tests (e.g., the Folded F -test). That is, the t -test assumes that the distributions of the two groups being compared are normal with equal variances. Although the authors of some statistics textbooks do not even mention the assumption of homogeneity of variance (e.g., Gravetter & Wallnau, 2011) as one required for the t -test, homoscedasticity is basic and necessary for hypothesis testing because the violations of this assumption “alter Type I error rates, especially when sample sizes are unequal” (Zimmerman, 2004, p. 173).

The preliminary test of the null hypothesis $\sigma_1^2 = \sigma_2^2$ versus the alternative $\sigma_1^2 \neq \sigma_2^2$ is conducted using the Folded F -test statistic $F = S_1^2 / S_2^2$. Common practice has been that if the Folded F -test is not statistically significant (e.g., $p \geq 0.05$), then the test of $\mu_1 = \mu_2$ versus $\mu_1 \neq \mu_2$ is calculated using the independent means t -test:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{S_{\bar{X}_1 \bar{X}_2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (1)$$

On the other hand, if the preliminary test is statistically significant (e.g., $p < 0.05$) and in addition there are unequal sample sizes, the independent means t -test should be avoided and the Satterthwaite's test should be used instead (Moser, Stevens, & Watts, 1989):

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad (2)$$

with

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}} \quad (3)$$

Recently, researchers have questioned the robustness of the conditional t -test with respect to Type I error and statistical power when the assumptions of normality and homoscedasticity are not met. Besides the unlikelihood of encountering real data that are normally distributed (Micceri, 1989), it is reported that there are also serious disadvantages of performing preliminary tests of variances equality (e.g., Moser et al., 1989; Zimmerman, 2004). Specifically, Moser et al. (1989) evaluated the impact of a preliminary variance test on the power and Type I error rate of the regular t -test and the Satterthwaite's test. Based on calculations of power and Type I error, they suggested that for equal sample sizes ($n_1 = n_2$), the t -test and the Satterthwaite's test had the same power and provided very stable Type I error rates close to the nominal alpha prescribed for the test of means. For unequal sample sizes ($n_1 \neq n_2$), the Satterthwaite's test still provided reasonable and stable Type I error rates close to the nominal significance level. In conclusion, Moser et al. recommended applying directly the Satterthwaite's test for testing the equality of means from two independent and normally distributed populations where the ratio of the variance is unknown.

Rasch, Kubinger, and Moder (2011) conducted a simulation study to compare the performance of the regular (Student) t -test, Welch test, and Wilcoxon U -test to investigate if we should perform the t -test conditionally after testing the assumptions. These authors suggested not testing the underlying assumptions of the t -test because such testing was not effective. Zimmerman (2004) found similar optimal results for the Welch-Satterthwaite separate-variance t -test if applied unconditionally whenever sample sizes were unequal and noted that the power of this test deteriorated if it was conditioned by a preliminary test. Grissom (2000)

PARAMETRIC TESTS FOR TWO POPULATION MEANS

argued that it is realistic to expect heteroscedasticity in data as well as outliers, and examined the effect of these factors on variance. He also addressed issues of robustness (i.e., control of Type I error rate) in the presence of heteroscedasticity and departures from normality, for which he suggested trimming as a way to stabilize variances.

Purpose of the Present Study

The purpose of this study was to explore the performance of the independent means *t*-test and two alternatives, Satterthwaite's test and the conditional *t*-test, by conducting a series of simulations under various manipulated conditions. The current study extended previous studies on the independent means *t*-test and its alternatives by taking into account the non-normality of population distribution and various levels of heteroscedasticity. Accordingly, extensive simulation conditions were included in this study: a wide range of total sample sizes (from 10 to 400 in contrast with 10 to 100 in [Rasch et al., 2011](#) and 30 and 60 in [Zimmerman, 2004](#)); various variance ratios between populations up to 1:20 (beyond the realistic maximum sample variance ratio of 1:12 suggested in [Grissom, 2000](#) and the great variance ratio of 1:16 mentioned in [Wilcox, 1987](#)); wide range of alpha set for testing treatment effects and testing homogeneity assumption for the conditional *t*-test; large range of non-normality (skewness from 0 to 6 and kurtosis from 0 to 25). In the study of [Rasch et al. \(2011\)](#), skewness and kurtosis were examined from 0 to 3 and 0 to 15, respectively. In addition, this paper provides some guidelines for researchers on the selection of an appropriate test for the equality of two population means.

Methodology

A simulation approach was used to explore and compare the behaviors of the independent means *t*-test, Satterthwaite's test, and the conditional *t*-test for two means because simulation allows for the controlling of designed factors.

Manipulated Factors

A crossed factorial mixed design included seven factors: (a) total sample size (10, 20, 50, 100, 200, 300, and 400), (b) sample size ratio between groups (1:1, 2:3, and 1:4), (c) variance ratio between populations (1:1, 1:2, 1:4, 1:8, 1:12, 1:16, and 1:20), (d) effect size for mean difference between populations ($\Delta = 0, .2, .5, \text{ and } .8$), (e) alpha set for testing treatment effects ($\alpha = 0.01, 0.05, 0.10, 0.15, 0.20, \text{ and } 0.25$),

(f) alpha set for testing homogeneity assumption for the conditional t -test ($\alpha = 0.01, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.40, 0.45,$ and 0.50), and (g) population distributions with varying kurtosis and skewness values (i.e., $\gamma_1 = 1.0$ and $\gamma_2 = 3.0$, $\gamma_1 = 1.5$ and $\gamma_2 = 5.0$, $\gamma_1 = 2.0$ and $\gamma_2 = 6.0$, $\gamma_1 = 0.0$ and $\gamma_2 = 25.0$, as well as $\gamma_1 = 0.0$ and $\gamma_2 = 0.0$ for the normal distribution, where γ_1 and γ_2 represent skewness and kurtosis, respectively). This crossed factorial design provided a total of 176,400 conditions for the conditional t -test and 17,640 conditions for the Satterthwaite's test and the regular t -test.

Data Generation

A random number generator, RANNOR in SAS/IML statistical software, was employed with a different seed value for each execution of the simulation program to generate data for this study. For each condition in the simulation, 100,000 samples were generated. The use of 100,000 replications provides a maximum standard error of an observed proportion (i.e., Type I error rate estimate) of .00158, and a 95% confidence interval no wider than $\pm .003$ (Robey & Barcikowski, 1992).

Statistically Analytical Procedures

For each sample generated, both the independent means t -test and Satterthwaite's test, each at a range of nominal alpha levels (i.e., 0.01 through 0.25), were conducted. The independent means t -test and Satterthwaite's test were investigated under a total of 17,640 conditions. In addition, the conditional t -test was conducted. The testing procedures for the conditional t -test were as follows. Firstly, the Folded F -test was implemented to examine the variance homogeneity assumption using a range of nominal alpha levels (i.e., 0.01 through 0.50). Based upon the results of the Folded F -test, either the independent means t -test or Satterthwaite's test was applied. Thus, for the conditional t -test, a total of 176,400 conditions were examined.

The simulation focused on Type I error rates and power. Type I error was examined when the population effect size (or two-group mean difference) was simulated zero; otherwise power was computed. Type I error rates were evaluated on the basis of the liberal criterion for robustness suggested by Bradley (1978). Given a nominal alpha level, Bradley's liberal criterion provides a plausible range of Type I error rates in which a test can be considered robust. The liberal criterion for the robustness is set at 0.5α around the nominal alpha. For example, when $\alpha = 0.05$, a test is considered robust when the Type I error rate falls between 0.025, which is given by $0.5 \cdot .05$, and .075, which is given by $1.5 \cdot .05$. When there was

PARAMETRIC TESTS FOR TWO POPULATION MEANS

considerable variability in the estimated Type I error and power across simulation conditions, eta-square analyses were conducted to examine the design factors related to the variability.

Results

The results of simulation are reported in the following order: (a) power of the Folded F -test for the test of equal variances, (b) Type I error control for the test of means, and (c) power for the test of means. Under the Type I error control for the test of means, an overview of Type I error rates, an analysis of design factors associated with Type I error control, and an analysis based on Bradley's liberal criterion for robustness are presented.

Power for the Folded F -Test

The distributions of statistical power estimates for the Folded F -test were examined across all conditions simulated in which population variances were not equal. As expected, when the alpha level used for the Folded F -test was small (e.g., 0.01 or 0.05), the average power was low. However, the power of the Folded F -test increased when the applied alpha level increased.

Nominal alpha levels of 0.05 and 0.25 for the Folded F -test were selected for further analysis of power. The average power of the Folded F -test based on simulation design factors is presented in Table 1. As seen in the table, the power remained stable regardless of distribution shapes; yet using the alpha level of 0.25 consistently yielded more power. The average powers for 0.05 and 0.25 alpha levels were around .81 and .90, respectively, across normal and non-normal distributions. Further, as the value of variance ratio increased, the power of the Folded F -test increased as well. Using the alpha level of 0.25 provided substantially more power when the variance ratios were small (i.e., variance ratio = 1:2 and 1:4). As the variance ratios increased, the power differences between the two nominal alpha levels decreased.

It is well-known that the power increases when the sample size increases. Using an alpha level of 0.05 for the Folded F -test yielded average power of .80 with sample size of 50 and of 0.90 with 100. In contrast, the average power reached .80 with as few as 20 observations and 0.90 with 50 observations using an alpha level of 0.25. The use of extremely unbalanced samples (sample size ratios of 1:4 or 4:1) reduced the power of the Folded F -test, but power advantages of the more liberal alpha level remained evident.

Table 1. The power of the Folded F -Test using $\alpha = .05$ and $\alpha = .25$

Condition	$\alpha = .05$	$\alpha = .25$	Condition	$\alpha = .05$	$\alpha = .25$
Total N			Variance ratio		
10	0.36	0.62	1:2	0.55	0.73
20	0.64	0.82	1:4	0.76	0.87
50	0.85	0.92	1:8	0.85	0.93
100	0.92	0.96	1:12	0.89	0.93
200	0.96	0.98	1:16	0.91	0.96
300	0.98	0.99	1:20	0.92	0.97
400	0.99	0.99			
N ratio			Shape		
1:04	0.74	0.86	0.0, 0	0.82	0.91
2:03	0.83	0.92	1.0, 3	0.81	0.90
1:01	0.85	0.93	1.5, 5	0.81	0.90
3:02	0.85	0.92	2.0, 6	0.81	0.89
4:01	0.80	0.82	0.0, 25	0.81	0.91

For Shape, the two values indicate skewness and kurtosis, respectively

Type I Error Control for the Test of Means

An overall view of the Type I error control of the tests is provided in Figures 1 and 2. These boxplots describe the distributions of the Type I error rate estimates under a nominal alpha level of 0.05 across all conditions in which the population means were identical. The first two plots are for the independent means t -test and Satterthwaite's test, respectively. The remaining plots delineate the Type I error rate estimates for the conditional t -test across the different conditioning rules (i.e., the alpha levels for the Folded F -test) that were investigated. For instance, the plot for C(01) provides the distribution of the Type I error rates for the conditional t -test when an alpha level of 0.01 was used with the Folded F -test as the rule to choose between the independent means t -test and Satterthwaite's test.

Note that in Figure 1 the independent means t -test has great dispersion of Type I error rates. In some conditions, this testing approach provides appropriate control of the Type I error probability while in others the Type I error rate is very different from the nominal alpha level. In contrast, Satterthwaite's approximate t -test provides adequate Type I error control in nearly all of the conditions simulated. The series of plots for the conditional t -test illustrate that the conditional test provides a notable improvement in Type I error control relative to the independent means t -test and the improvement increases as the alpha level for the Folded F -test increases. This improvement occurs because the statistical power of the Folded F -test increases as the alpha level increases. That is, the ability of the Folded F -test

PARAMETRIC TESTS FOR TWO POPULATION MEANS

to detect variance heterogeneity (and to subsequently steer us away from the independent means t -test and steer us to Satterthwaite's test) increases with the alpha level for this test, which supports the argument of insufficient power when using a more conservative alpha level for a preliminary analysis.

Considering that the power of Folded F -test was substantially lower when the total sample size was 10 or 20 (see Table 1) and the behavior of conditional t -test heavily depended on the power of the Folded F -test, we inspected the Type I error rates of the conditional t -test only for total sample size greater than 20. As speculated, the Type I error rates of the conditional t -test across different alpha levels are almost identical to that of Satterthwaite's test if the decision of conditional t -test is made at $\alpha = 0.10$ or greater (see Figure 2).

Figure 1. Distributions of estimated Type I error rates for independent means t -test (t -test), Satterthwaite's test, and conditional t -test ($\alpha = 0.05$) for all sample size conditions. C(01) denotes the conditional t -test at $\alpha = 0.01$ of the Folded F -test

Figure 2. Distributions of estimated Type I error rates for independent means t -test (t -test), Satterthwaite's test, and conditional t -test ($\alpha = 0.05$) for $N > 20$. C(01) denotes the conditional t -test at $\alpha = 0.01$ of the Folded F -test

Impact of Simulation Design Factors on Type I Error Controls

Variance heterogeneity. The large dispersion of Type I error rates for the independent means t -test resulted in large part from the variance heterogeneity that was included in the simulation conditions. Figure 3 presents the distributions of Type I error rates for the independent means t -test with the results disaggregated by population variance ratio. Note that as the population variance ratio increases, both the average Type I error rate and the dispersion of Type I error rates increase. On the other hand, both Satterthwaite's test and the conditional t -test provide good control of Type I error rate even if the population variances in the two groups are heterogeneous (Figure 3).

Of course, the independent means t -test is known to be relatively robust to violations of the assumption of variance homogeneity if the sample sizes in the two groups are equal. This phenomenon is illustrated in Figure 4. Note that the Type I error rate for the independent means t -test is maintained near the nominal 0.05 level if sample sizes are equal. With disparate sample sizes in the two groups, the independent means t -test either becomes conservative (Type I error rates lower than the nominal alpha level) or liberal (Type I error rates higher than the nominal level) depending upon the relationship between sample size and population variance. In contrast, both Satterthwaite's test and the conditional t -test evidence much

PARAMETRIC TESTS FOR TWO POPULATION MEANS

Figure 3. Distributions of estimated Type I error rates by variance ratio (1:1, 2:1, 4:1, 8:1, 12:1, 16:1, 20:1) at $\alpha = 0.05$.

Figure 4. Distributions of estimated Type I error rates by sample size ratio (1:4, 1:3, 1:1, 3:1, 4:1) at $\alpha = 0.05$.

improved Type I error control under variance heterogeneity when samples sizes are unequal.

Other design factors related to Type I error control. The variability in the estimated Type I error rates for the three tests of means was analyzed by computing the eta-squared value associated with each simulation design factor and the first-order interactions. For the independent means *t*-test, the factors associated with variability in estimated Type I error rates were sample size ratio ($\eta^2 = 0.69$) and the interaction between sample size ratio and variance ratio ($\eta^2 = 0.22$). For Satterthwaite's test and the conditional *t*-test respectively, the major factors were sample size ratio ($\eta^2 = 0.15$; $\eta^2 = 0.18$), total sample size ($\eta^2 = 0.18$; $\eta^2 = 0.14$), and the interaction between sample size ratio and total sample size ($\eta^2 = 0.26$; $\eta^2 = 0.36$). An analysis of the sole impact of distribution shape on Type I error rates of the three tests showed that Type I error rate of Satterthwaite's test was most affected ($\eta^2 = 0.07$). While Type I error rate of the independent means *t*-test was least impacted by the distribution shape ($\eta^2 = 0.001$), Type I error rate of the conditional *t*-test was also impacted ($\eta^2 = 0.04$) but to a much lesser degree in comparison with that of Satterthwaite's test.

The mean Type I error rates by total sample size and distribution shape for the independent means *t*-test and Satterthwaite's test under the nominal alpha level of .05 are presented in Figures 5 and 6. The graph for the conditional *t*-test is similar to that for Satterthwaite's test. The mean Type I error rates of the independent means *t*-test are much above the nominal alpha level in all conditions of distribution shapes and total sample sizes (see Figure 5). Although the mean estimated Type I error rates decrease with larger samples, they remain substantially greater than 0.05. In contrast, both Satterthwaite's test and the conditional *t*-test provided much better Type I error control except for extremely small sample sizes (i.e., total sample size of 10 or 20) or the extremely skewed distribution (e.g., skewness of 2) (see Figure 6 for Satterthwaite's test).

The factors related to the Bradley proportions vary across tests (Table 2). For the independent means *t*-test, sample size ratio and variance ratio between the two populations emerged as primary factors making an impact on the Type I error control. Although the overall proportions of cases meeting the Bradley's criterion for the independent means *t*-test were very low (below 50%), the Type I error rates were perfectly controlled when the homogeneity of variance assumption was met (i.e., variance ratio between groups = 1:1). As the disproportion of two group variances became larger to 1:20, the Type I error control of independent means *t*-test diminished considerably. When the two groups have equal sample size, the

PARAMETRIC TESTS FOR TWO POPULATION MEANS

independent means t -test adequately controlled the Type I error rates within the Bradley's criterion for 91% of the conditions. The imbalance of sample size between groups worsened the Type I error control noticeably. On the other hand, the adequacy of Type I error control of the independent means t -test appears independent of total sample size and the shape of distribution. That is, the proportions meeting the Bradley's criterion were consistently low irrespective of total sample size and distribution shape.

Figure 5. Mean Type I error rate by total sample size for the independent means t -test

Figure 6. Mean Type I error rate by total sample size for the Satterthwaite's test

Table 2. Proportions of cases meeting the Bradley's Liberal Criterion at $\alpha = 0.05$

Condition	t-test	Conditional	Satterthwaite	Condition	t-test	Conditional	Satterthwaite
Total <i>N</i>				Variance ratio			
10	0.45	0.68	0.65	1:1	1.00	0.94	0.92
20	0.49	0.76	0.82	1:2	0.62	0.93	0.95
50	0.45	0.93	0.95	1:4	0.40	0.91	0.93
100	0.43	0.97	0.97	1:8	0.29	0.90	0.91
200	0.42	1.00	1.00	1:12	0.27	0.89	0.89
300	0.41	1.00	1.00	1:16	0.25	0.89	0.89
400	0.41	1.00	1.00	1:20	0.23	0.89	0.90
<i>N</i> ratio				Shape			
1:4	0.18	0.98	0.97	0.0, 0	0.43	0.96	0.97
2:3	0.67	0.98	0.98	1.0, 3	0.44	0.96	0.97
1:1	0.91	0.97	0.97	1.5, 5	0.44	0.93	0.94
3:2	0.28	0.91	0.91	2.0, 6	0.46	0.77	0.78
4:1	0.14	0.70	0.74	0.0, 25	0.41	0.91	0.91

Conditional indicates conditional *t*-test at $\alpha = .25$ of Folded *F*-test. Shape values indicate skewness and kurtosis, respectively.

The impact of variance ratio and sample size ratio on the Type I error control appears minimal for the Satterthwaite's test and the conditional *t*-test. Both tests showed adequate levels of Type I error control in the majority of conditions regardless of variance ratio and sample size ratio. Instead, total sample size and the skewness of the distribution were associated with the Bradley proportions of both Satterthwaite's test and the conditional *t*-test. When the total sample size was 10, the proportions meeting the criterion dropped to about 65%. In this total sample size condition ($N = 10$), the conditional *t*-test showed slightly better control of Type I error (68%) than the Satterthwaite's test. Interestingly, for both tests the proportions meeting the Bradley's criterion were affected by skewness but not by kurtosis (see Table 2).

Statistical Power Analysis

Although Satterthwaite's test generally provides superior Type I error control, it is not always the best test to select because of the potential for power differences. When the assumptions are met, the independent means *t*-test is the most powerful test for mean differences. For this simulation study, power comparisons were made only for conditions in which both Satterthwaite's test and the conditional *t*-test procedures evidenced adequate Type I error control by Bradley's (1978) benchmark. The distributions of power estimates (at a nominal alpha level of 0.05

PARAMETRIC TESTS FOR TWO POPULATION MEANS

for the tests of means) for Satterthwaite's test and the conditional t -test (using an alpha level of 0.25 for the Folded F -test) showed that the power differences between the tests were small.

Figure 7 presents a scatter plot of the power estimates for Satterthwaite's test and the conditional t -test (using an alpha level of 0.25 for the Folded F -test). Data points above the line represent conditions in which the conditional t -test was more powerful than Satterthwaite's test, while those below the line are conditions in which Satterthwaite's test is more powerful. Overall, the conditional t -test, using an alpha level of 0.25 for the Folded F -test of variances, was more powerful in 29% of the conditions, while Satterthwaite's test was more powerful in only 23% of the conditions (identical power estimates were obtained in the other conditions).

To identify research design factors associated with power differences between these two tests, the percentages of conditions in which each test evidenced power advantages were disaggregated by the simulation design factors (Table 3). For conditions with homogeneous variances, the conditional t -test evidenced more power than the Satterthwaite's test in 61.64% of the conditions, while the Satterthwaite's test was more powerful in 20.55% of the conditions (in the

Figure 7. Scatterplot of power estimates for the conditional t -test and Satterthwaite's approximate t -test

remaining conditions, the two tests evidenced equal power). As the variance ratios increased, the power advantages of the conditional t -test diminished, such that the Satterthwaite's test was more often the more powerful test when the population variance ratio was 1:8 or larger.

With balanced samples the conditional t -test was more powerful in 35.38% of the conditions and Satterthwaite's test was never more powerful. With unbalanced samples in which the larger sample is drawn from the population with the larger variance (in heterogeneous populations), the Satterthwaite's test presents notable power advantages (44.77% and 51.03% of the cases with sample size ratios of 2:3 and 1:4, respectively). In contrast, when the larger sample is drawn from the population with the smaller variance, the conditional t -test evidences more power than the Satterthwaite's test (47.68% and 48.52% of the conditions with sample size ratios of 3:2 and 4:1, respectively). The results by total sample size show that the conditional t -test is more powerful in more conditions, except for the smallest sample sizes examined ($N = 10$). Finally, the conditional t -test is more powerful in more conditions for all distribution shapes except for the most skewed distribution examined (i.e. skewness of 2, kurtosis of 6).

Table 3. Percentage of simulation conditions by simulation design factors in which the conditional t -test and Satterthwaite's test were more powerful

Condition	Conditional	Satterthwaite	Condition	Conditional	Satterthwaite
Total N			Variance ratio		
10	48.97	51.03	1:1	61.64	20.55
20	54.89	44.11	1:2	43.15	34.97
50	44.79	28.83	1:4	31.03	29.35
100	29.22	20.98	1:8	21.02	22.08
200	19.43	12.38	1:12	16.56	18.06
300	11.62	9.52	1:16	14.62	17.63
400	8.19	7.81	1:20	13.25	15.81
N ratio			Shape		
1:4	10.59	53.95	0.0, 0	29.20	15.67
2:3	9.21	44.77	1.0, 3	29.60	21.98
1:1	35.58	0.00	1.5, 5	28.17	24.63
3:2	47.68	4.04	2.0, 6	21.69	24.51
4:1	48.52	4.54	0.0, 25	34.98	27.50

Conditional indicates conditional t -test at $\alpha = 0.25$ of Folded F -test. Shape values indicate skewness and kurtosis, respectively

Conclusion

The testing of differences between two population means is a fundamental statistical application, but controversy about the appropriate test to use has been evident for many years. When conducting the independent means t -test, major statistical software programs (e.g., SAS and SPSS) automatically produce the results of the independent means t -test and the alternative Satterthwaite's test. Depending on the statistical significance of homogeneous variance testing (Folded F -test in SAS and Levene's F test in SPSS), researchers are recommended to follow one of the options; this has been a common practice in studies comparing two population means. However, recent studies on the conditional t -tests in comparison to Satterthwaite's test have strongly supported the Satterthwaite's test over the conditional t -test and suggested even abandoning the conventional practice of selecting one of the options based on the results of the homogeneity of variance test. Considering the ongoing controversy surrounding these tests and the frequency with which two means are compared in applied research, in this simulation study we investigated the performance of the independent means t -test, Satterthwaite's approximate t -test, and the conditional t -test under the manipulated conditions of population distribution shape, total sample size, sample size ratio between groups, variance ratio between populations, difference in means between populations, alpha level for testing the treatment effect, and alpha level for testing the homogeneity assumption for the conditional t -test. Type I error control and power analysis were used to examine the performance of these testing procedures.

As expected, the independent means t -test performed very well on Type I error control when the homogeneity assumption was met regardless of the tenability of the normality assumption. This reminds us of the long-known property that the independent means t -test requires the homogeneity assumption to be met and this test is robust to violations of the normality assumption when two population variances are equal. Furthermore, the independent means t -test showed adequate Type I error control when sample sizes in the two groups were equal under the normal distribution. This re-emphasizes another well-known property that the independent means t -test is robust to violations of the homogeneity assumption when the sample sizes are equal under the normal distribution. Under these conditions, the independent means t -test is the best method to test the difference between two independent means. This testing procedure also provides more statistical power. On the other hand, the t -test evidenced poor Type I error control under heterogeneous variances with non-normal distributions. Thus, two

alternatives, Satterthwaite's test and the conditional t -test, were considered in this study.

It was also found that the Type I error rate of the conditional t -test was affected by the alpha level for the Folded F -test that was used to test the homogeneity assumption of population variances. The more conservative alpha levels for the Folded F -test resulted in larger Type I error rates for the conditional test because of lower statistical power, such that the Folded F -test is less likely to detect the true difference between population variances. This leads us to re-consider the conventional procedures for examining the difference between two population means. Thus, the conditional t -test using a relatively large alpha level for the Folded F -test may be an appropriate alternative.

Overall, Satterthwaite's test performed best in control of Type I error rate but the conditional t -test also yielded comparable results using a large alpha level of .25 for the Folded F -test. Both alternatives made a tremendous improvement in Type I error control, compared to the independent means t -test, when group variances were unequal. Extreme skewness (e.g. skewness of 2) contaminated the Type I error control for both alternative testing procedures. Kurtosis seemed not to have this kind of impact. Increasing total sample size was found in this study to improve Type I error control for both testing procedures, but not for the independent t -test. When total sample size was 200 or more, Bradley's rates were 100% for both alternative testing procedures. Although Satterthwaite's test provides slightly better Type I error control, the use of the conditional t -test may have a slight power advantage.

Recommendations. With equal sample size the independent means t -test is the appropriate testing procedure to examine the difference of two independent group means because it provides adequate Type I error control and more statistical power. With unequal sample size the Folded F -test can provide reasonable guidance in the choice between the independent t -test and Satterthwaite's test. A large alpha level of .25 is recommended to evaluate the results of the Folded F -test. If the F value is not statistically significant at this large alpha level, then the independent means t -test should be used. In contrast, if the F value is statistically significant at this large alpha level, then Satterthwaite's test should be chosen. Finally, the confidence in this conditional testing procedure increases as the sample sizes become larger. To adequately control for Type I error rate in the conditional testing procedure, a total sample size of at least 200 is recommended with extremely skewed populations (e.g. skewness of 2). For less skewed populations, a total sample size of at least 100 is recommended. With a total sample size smaller than these

PARAMETRIC TESTS FOR TWO POPULATION MEANS

recommended in the corresponding conditions, the Type I error control resulting from any of these testing procedures may be questionable.

References

- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)
- Bridge, P. D., & Sawilowsky, S. S. (1999). Increasing physicians' awareness of the impact of statistics on research outcomes: Comparative power of the t-test and Wilcoxon Rank-Sum test in small samples applied research. *Journal of Clinical Epidemiology*, 52(3), pp. 229-235. doi: [10.1016/S0895-4356\(98\)00168-1](https://doi.org/10.1016/S0895-4356(98)00168-1)
- Cody, R. P., & Smith, J. K. (1997). *Applied statistics and the SAS programming language* (4th ed.). Upper Saddle River, NJ: Prentice Hall.
- Gravetter, F. J., & Wallnau, L. B. (2011). *Essentials of Statistics for the Behavioral Sciences* (6th ed.). Belmont, CA: Wadsworth, Cengage Learning.
- Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology*, 68(1), 155-165. doi: [10.1037/0022-006X.68.1.155](https://doi.org/10.1037/0022-006X.68.1.155)
- Hayes, A. F., & Cai, L. (2007). Further evaluating the conditional decision rule for comparing independent means. *British Journal of Mathematical and Statistical Psychology*, 60(2), 217-244. doi: [10.1348/000711005X62576](https://doi.org/10.1348/000711005X62576)
- Micceri, T. (1989). The unicorn, the normal curve, and the other improbable creatures. *Psychological Bulletin*, 105(1), 156-166. doi: [10.1037/0033-2909.105.1.156](https://doi.org/10.1037/0033-2909.105.1.156)
- Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t test versus Satterthwaite's approximate f test. *Communications in Statistics: Theory and Methods*, 18(11), 3963-3975. doi: [10.1080/03610928908830135](https://doi.org/10.1080/03610928908830135)
- Rasch, D., Kubinger, K. D., & Moder, K. (2011). The two-sample t test: pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1), 219-231. doi: [10.1007/s00362-009-0224-x](https://doi.org/10.1007/s00362-009-0224-x)
- Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45(2), 283-288. doi: [10.1111/j.2044-8317.1992.tb00993.x](https://doi.org/10.1111/j.2044-8317.1992.tb00993.x)
- Schlotzhauer, S., & Littell, R. (1997). *SAS System for elementary statistical analysis* (2nd ed.). Cary, NC: SAS Institute.

Wilcox, R. R. (1987). New designs in analysis of variance. *Annual Review of Psychology*, 38, 29-60. doi: [10.1146/annurev.ps.38.020187.000333](https://doi.org/10.1146/annurev.ps.38.020187.000333)

Zimmerman, D. W. (2004). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology*, 57(1), 173-181. doi: [10.1348/000711004849222](https://doi.org/10.1348/000711004849222)

Multiple Imputation When Rate of Change is the Outcome of Interest

Manisha Desai
Stanford University
Stanford, CA

Aya A. Mitani
Boston University
Boston, MA

Susan W. Bryson
Stanford University
Stanford, CA

Thomas Robinson
Stanford University
Stanford, CA

Little research has been devoted to multiple imputation (MI) of derived variables. This study investigates various MI approaches for the outcome, rate of change, when the analysis model is a two-stage linear regression. Simulations showed that competitive approaches depended on the missing data mechanism and presence of auxiliary terms.

Keywords: Multiple imputation, derived variables, active imputation, passive imputation, mixed effects model, two-stage linear regression, slope

---

Introduction

Multiple imputation (MI) techniques, although sound in theory, often present challenges in practice. One issue involves how best to multiply impute derived variables. These include interaction terms, higher-order terms, variables that are functions of others such as body mass index (BMI) – a function of height and weight – and rate of change in a measure (or the slope). Consider interaction terms. A natural question is whether the main effects should first be imputed so that the interaction term can then be derived, or whether the interaction terms should be imputed as any other variable. The former approach falls under the umbrella of what is called passive imputation methods and the latter under the umbrella of active imputation techniques (von Hippel, 2009; van Buuren, 2012). Similar issues arise in other contexts. For example, if one has data on height but not on weight for a subject, where BMI is of interest, a passive approach would be to first impute weight and then derive BMI, while an active approach would impute BMI directly as any other variable along with weight.

This work is motivated by the ongoing Stanford GOALS randomized clinical trial, which is designed to evaluate a novel intervention in reducing BMI

Dr. Desai is Professor of Biomedical Informatics Research in the Department of Medicine. Email her at manishad@stanford.edu. Aya Mitani is a graduate student in the Department of Biostatistics. Susan Bryson is a staff scientist in the Department of Pediatrics. Dr. Robinson is Professor in the Departments of Pediatrics and Medicine.

(weight in kilograms divided by the square of height in meters) among overweight and obese children (Robinson et al., 2013). The primary outcome of the trial is the rate of change in BMI, which entails measuring BMI at 4 time points over a 3-year period (at baseline and at approximately 1, 2, and 3 years post-randomization). As typically occurs in clinical trials, it is anticipated that a proportion of children will be missing one or more follow-up BMI measurements upon which to calculate the rate of change. The focus of this study is on how to impute slopes when slope is the outcome of interest and missing for a proportion of subjects due to incomplete BMI measurements.

Multiple imputation (MI)

MI is a simulation-based method for handling missing data (Rubin, 1996). There are three main steps involved in its use. The first is the imputation step where plausible values for missing data are drawn and filled in to create a full data set. This is done multiple, or m , times. The second step is the model-fitting step where the scientific model is fitted to each of the m data sets. The third step is the combination step where Rubin's rules are applied to provide one summary result (Little & Rubin, 2002). Standard MI produces results that are statistically valid when the data are missing at random (MAR) (Ibid.). Essentially this means that missingness can be related to observed features, although conditional on such features, missingness is not related to unobserved data. The data are considered missing completely at random (MCAR) if missingness is unrelated to both unobserved and observed features, and the data are not missing at random (NMAR) if missingness is related to unobserved features. See Rubin's seminal text on missing data for a complete description on mechanisms of missingness (Ibid.). MI may also be applied when the data are NMAR, but this requires explicit modeling of the missing data mechanism (Ibid.). Further, even when the data are NMAR, in the presence of sufficiently strong auxiliary data, applying MI assuming MAR may be reasonable (Collins et al., 2001). Importantly, MI yields estimates with desirable statistical properties under the same conditions in which the more ideal maximum likelihood-based methods for handling missing data produce estimates (Little & Rubin, 2002). MI, in contrast, is more accessible as it is available in several mainstream packages, while software is not available for all cases when applying maximum likelihood-based approaches.

Specification of the imputation model (Step 1 of the MI process described above) is a critical step when applying MI, and requires the user to make key decisions. This includes the choice of one of two main approaches for imputing

the data. These are the joint modeling (JM) approach and the fully conditional specification (FCS) approach ([van Buuren, 2007](#)). Briefly, the JM approach involves specification of a joint model for the data from which the posterior predictive distribution for the missing data can be derived in order to impute the missing data. The approach typically assumes a multivariate normal distribution (although other parametric distributions are possible) and has well established and tractable statistical properties (*Ibid.*). FCS was developed to accommodate data sets of mixed type (i.e., when categorical variables are present), and therefore is useful when a plausible multivariate distribution may not be appropriate. Although the statistical properties are not tractable, use of FCS has been shown to be comparable to that of JM in practice and through simulation studies (*Ibid.*). In his excellent text on MI, van Buuren (2012) described other considerations, including which variables to include as predictors, the number of imputations to be performed, and how to handle derived variables.

Current literature on active versus passive imputation approaches

There are two main umbrellas of approaches that can be used to impute derived variables: active and passive. In general, active imputation involves imputing the derived variable as just another variable or JAV method ([Seaman et al., 2012](#)). In contrast, purely passive imputation – as implemented in STATA through ICE, which was developed by Patrick Royston (and more recently through MI IMPUTE CHAIN) or MICE in R – is an iterative process that involves using the derived variable in the imputation of all variables with the exception of the main effects ([van Buuren, 2012](#); [Royston, 2009](#)). The derived term is then created once the main effects have been imputed (*Ibid.*). The advantage of an active approach is that it adheres to the generally well-accepted rule that the relationships specified in the scientific model are included in the imputation model (i.e., that the imputation model is congenial with the scientific model leading to what is called proper imputations ([van Buuren, 2012](#); [Rubin, 1996](#); [Rubin, 1987](#); [Meng, 1994](#)). For a more detailed discussion on this topic see [Rubin \(1987\)](#) and [Meng \(1994\)](#). This ensures that interrelationships are preserved appropriately, leading to unbiased descriptions of the relationship of interest. Passive approaches, on the other hand, ensure that imputed values adhere to consistent relationships and are therefore plausible. For example, under active imputation, if BMI is imputed as just another variable, it is possible to get a value of say 30 instead of 20 when observed weight is 54.5 kg (120 lbs.) and imputed height is 1.65 meters (65 inches). Passive imputation can prevent such inconsistencies.

Research on optimal approaches for imputing derived variables has been limited. Much of the literature has focused on one of the more common derived variables – interaction terms (von Hippel, 2009; Seaman et al., 2012; Allison, 2002). Allison (2002) briefly introduced the problem of multiply imputing interaction terms in his widely used and practical textbook on missing data. von Hippel (2009) discussed two major approaches for handling missing data of interaction and higher-order terms under an assumption that the data are MCAR – a “transform then impute” or active approach and an “impute then transform” or passive-like approach, and provided a mathematical argument to demonstrate that under MCAR the active approach is less biased than the passive. von Hippel (2009) also argued that the purely passive approach is similarly biased. White et al. (2011) compared passive and active approaches for interaction terms under MAR and through simulations demonstrated that they are both biased but comparable. The authors recommend sensitivity analyses of both approaches in assessing robustness of findings. Mitani and others (2015) evaluated active and passive methods across different software platforms for interaction terms that involved categorical variables. They found an improved passive approach – where an interaction between the outcome and main effect is included in the imputation of the other main effect is included in a passive imputation -- outperformed both active and purely passive approaches. van Buuren (2012) emphasizes the importance of internally consistent imputed values, and therefore recommends passive approaches that are designed for this purpose. Although active approaches may yield valid statistical inferences under certain conditions, he argues that if implausible combinations of imputed values are produced that would not have been observed otherwise it undermines the imputation process. van Buuren (2012) discussed active versus passive imputation in the context of interaction terms, sum scores, and other derived variables such as ratios. However, their variability in performance for measures of rate of change or slope was not discussed. Although imputation in the context of mixed effects models has received some attention (e.g., Schafer, 1997b), no research has been conducted on imputation when the scientific model is a two-stage linear regression. Our paper focuses on situations in which slope – a variable derived from repeated measurements – is the outcome of interest and the scientific model is a two-stage linear regression. Furthermore we consider a wider range of active and passive methods under a variety of missing data mechanisms that include the NMAR condition.

The Stanford GOALS trial

This work is motivated by a large-scale community-based randomized clinical trial currently being conducted at Stanford called Stanford GOALS (Robinson et al., 2013). The objective of the trial is to evaluate a novel multi-level, multi-component intervention to treat overweight and obese children aged 7-11 years old. The intervention consists of access to after-school team sports, home visits for behavioral counseling and modifications of the home environment, and primary care counseling. Two-hundred forty subjects are randomized to 1 of 2 groups – the intervention group or a control group receiving health and nutrition education plus their current standard of care. The primary outcome is a derived measure – the rate of change in BMI. Baseline BMI is measured for all randomized subjects. BMI is then measured at approximately 1, 2, and 3 years post randomization, and the outcome is a function of the 4 BMI measurements. As with most prospective randomized clinical trials, some subjects are expected to be missing one or more of the follow-up BMI measures.

The data are not expected to be missing completely at random. Missingness may be related to follow-up time, treatment arm, baseline BMI and even rate of change in BMI itself. However, conditional on baseline and intermittent BMI values, as well as a number of related measures including triceps skinfold, waist circumference, blood pressure, fasting blood lipids, insulin, glucose, and hemoglobin A1c to name a few, recorded at baseline and annually, we expect the data to be MAR. Our analytic plan specifies that we will use MI to impute the outcome or rate of change. For subjects who have at least 2 measurements, slope can be calculated with the baseline and subsequent measurements. Thus, the plan is to impute slope for those subjects who only have a baseline BMI and no other follow-up measurements.

Slope model versus mixed effects models

A simplified version of the primary analysis proposed for Stanford GOALS is a two-stage linear model or what we will refer to here as a slope model. More specifically, in the first stage, we derive a slope for all individuals with at least two BMI measurements by regressing BMI on time for each individual:

$$BMI_{ij} = \alpha_{0i} + \alpha_{1i} time_{ij} + \varepsilon_{ij}, \quad (1a)$$

where $time_{ij}$ represents time in months since randomization for the i^{th} person at the j^{th} visit, and ε_{ij} represents normally distributed random error; and in the second, we regress slope on treatment arm:

$$\hat{\alpha}_{1i} = \eta_0 + \eta_1 Treatment_i + \delta_i, \quad (1b)$$

where $Treatment_i$ is an indicator for whether the i^{th} subject is in the intervention group, and δ_i represents normally distributed random error. As a secondary analysis, we will fit a mixed effects model where we regress BMI on treatment arm, time, and an interaction between time and treatment, and include subject-specific random effects for the intercept and slope with time:

$$BMI_{ij} = (\gamma_{0i} + \beta_0) + \beta_1 Treatment_i + (\beta_2 + \gamma_{1i})time_{ij} + \beta_3 Treatment_i \times time_{ij} + \xi_{ij}, \quad (2)$$

where γ_{0i} and γ_{1i} are normally-distributed subject-specific random effects – a random intercept and slope for time for the i^{th} subject – and ξ_{ij} is the normally distributed error term for the j^{th} observation corresponding to the i^{th} subject. The parameter of interest is the difference in rate of change in BMI between the treated and non-treated groups. For the slope model, this is η_1 and for the mixed effects model, this is β_3 . Although equivalent in that both models assume subjects have their own random intercept and random slope, the procedures for fitting these models differ and may yield different estimates particularly in the presence of missing data. The mixed effects model with random intercepts and random slopes or RIRS model – yields an estimate based on any subject who contributes at least one outcome of BMI and relies on an assumption that the data are missing at random. The slope model yields an estimate based on any subject who contributes at least two outcomes of BMI (at baseline and at least one follow-up time point). Thus, those who contribute only a baseline measurement are excluded from the analysis. Incorporating MI methods into the process to impute rate of change for those missing all 3 follow-up BMI values and then applying the slope model relies on a MAR assumption about the missing data as well.

Issues with choosing a two-stage linear regression or slope model over the RIRS model have been discussed at great length (e.g., Rogosa & Saner, 1995; Gelman, 2005; Jusko & Shively, 2013; Duch & Stevenson, 2005). There are several reasons, however, why the slope model might be chosen over the RIRS model (Rogosa & Saner, 1995; Gelman, 2005; Liao et al., 2013). Specialized software is required in the latter case and the complexity of the model parameterization can compromise interpretability; in contrast, Equation 1b above

is easy to interpret compared to a mixed effects model that in a longitudinal setting involves interpreting an interaction term. In addition, in the process of fitting the slope model, insight into the data can be gained; in particular, heterogeneity in estimates of intercepts and slopes across subjects can be assessed. These features are not typically examined when fitting the RIRS model; rather, they are treated as nuisance parameters. There may be advantages, however, in the properties of estimates resulting from RIRS models over those resulting from the two-stage models. Gelman (2005) advocated the two-stage model over the mixed effects model in certain conditions, because of the simplicity and insight gained which are strong reasons to choose the two-stage model when it is appropriate. It was noted there are situations where the advantages of the mixed effects model outweigh its complexity – e.g., in the longitudinal setting, if data within individuals are sufficiently sparse, it may be desirable to borrow strength from other individuals by fitting the RIRS model.

Therefore, the goal is to determine whether different approaches to imputing slope – when a two-stage linear model or slope model is of interest – affect properties of estimates of interest. In particular we are interested in approaches that can be employed by mainstream packages that have incorporated MI procedures. To that end, variability will be evaluated among active versus passive methods of imputation that are readily accessible in STATA and R. We focus on situations where the treatment effect is of interest and where auxiliary information may or may not be present.

Methodology

A simulation study was conducted to closely mimic the Stanford GOALS trial, where there were two treatment arms with 120 subjects per group. Other parameters were not based on empirical findings, as the study is currently ongoing. Specifically, we assumed there was a treatment effect such that those in the treatment group did not increase their BMI over time, while those in the control group increased by 0.5 BMI units per year, the percent of observations (not individuals) missing varied (20% and 40%), the true model varied (RI or slope/RIRS model), and auxiliary information was and was not present (Table 1).

Table 1. Description of Scenarios

Scenario	Model for Data Generation	Model for Missingness	Average % Missing	Auxiliary Data Available	Auxiliary Data Related to Variables	Auxiliary Data Related to Missingness
A1	RI model	MAR	40	No	N/A	N/A
A2	RI model	NMAR	40	No	N/A	N/A
B3	RI model	MAR	40	Yes	Yes	No
B4	RI model	MAR	40	Yes	Yes	Yes
B5	RI model	NMAR	40	Yes	Yes	No
B6	RI model	NMAR	40	Yes	Yes	Yes
B7	RI model	NMAR	20	Yes	Yes	Yes
C8	RIRS/Slope model	MAR	40	Yes	Yes	Yes
C9	RIRS/Slope model	NMAR	40	Yes	Yes	Yes

Generating BMI

The true model from which the data were generated under Scenarios A and B was a mixed effects model with random intercept only – which we will refer to as the RI model – a special case of the RIRS model which includes both random intercept and random slope for time for each subject (Table 1). The RI model can be expressed as:

$$BMI_{ij} = (\gamma_{0i} + \beta_0) + \beta_1 Treatment_i + \beta_2 time_{ij} + \beta_3 Treatment_i \times time_{ij} + \zeta_{ij},$$

where $\beta_3 = -0.5$, $\zeta_{ij} \sim N(0,1)$, $\gamma_{0i} \sim N(0,2)$.

The true model from which the data were generated under Scenario C was the slope model, so that we first generated slopes from the following model:

$$\alpha_{1i} = \eta_0 + \eta_1 Treatment_i + \delta_i$$

where $\eta_1 = -0.5$, $\delta_i \sim N(0,0.7)$. BMI values for each subject were then generated based on the realized slope values and random intercepts according to the following model:

$$BMI_{ij} = \alpha_{0i} + \hat{\alpha}_{1i} time_{ij} + \varepsilon_{ij}$$

where $\alpha_{0i} \sim N(25,2)$, $\varepsilon_{ij} \sim N(0,1)$, and $time_{ij}$ takes on values of $j=0, 1, 2$, or 3 to represent the j^{th} visit of the i^{th} individual. Note that this is also the RIRS model, as

MULTIPLE IMPUTATION FOR RATE OF CHANGE

each individual has its own intercept and slope generated as random variables from a normal distribution.

Generating auxiliary terms

Two auxiliary terms – one continuous to represent age, and one dichotomous to represent gender – were generated. In some scenarios (B3 and B5), the auxiliary variables were related to baseline BMI and the rate of change in BMI or the slope, and in others they were additionally related to missingness (B4, B6, B7, C8, C9) (Table 1). Specifically, in Scenarios B3-B7 and C8-C9, one continuous variable representing *Age*, was generated for the i^{th} person as follows:

$$Age_i = 5 \times Slope_i + 2 \times BaselineBMI_i + \varepsilon_{ij}$$

where the error was normally distributed with variance of 4. An indicator for male gender was generated as follows:

$$\text{logit}(Male_i) = -5 + 0.5 \times Slope_i + 0.2 \times BaselineBMI_i$$

Inducing missingness

Missingness (under MAR) was related to time since randomization and treatment arm for Scenarios A1 and B3:

$$\begin{aligned} \text{logit}(missing_{ij}|Treatment_i = control) &= -1.5 + 2 \times time_{ij}, \text{ and} \\ \text{logit}(missing_{ij}|Treatment_i = intervention) &= 2. \end{aligned}$$

Missingness (under MAR) was additionally related to auxiliary terms for Scenarios B4 and C8:

$$\begin{aligned} \text{logit}(missing_{ij}|Treatment_i = control) &= -3 + 2 \times time_{ij} + 1.5 \times I[Age_i < 50] \\ &\quad + I[Male_i = 1], \text{ and} \\ \text{logit}(missing_{ij}|Treatment_i = intervention) &= -1.8 + I[Male_i = 0]. \end{aligned}$$

Missingness (under NMAR) was related to time since randomization, treatment arm, and slope for Scenarios A2 and B5:

$$\begin{aligned} \text{logit}(\text{missing}_{ij} | \text{Treatment}_i = \text{control}) &= -1.5 + 2 \times I[\text{slope}_i \geq 0.2] + \text{time}_{ij}, \text{ and} \\ \text{logit}(\text{missing}_{ij} | \text{Treatment}_i = \text{intervention}) &= -2.5 + I[\text{slope}_i < 0.5]. \end{aligned}$$

Missingness (under NMAR) was additionally related to auxiliary terms for Scenarios B6, B7, and C9:

$$\begin{aligned} \text{logit}(\text{missing}_{ij} | \text{Treatment}_i = \text{control}) &= -2.5 + 2 \times I[\text{slope}_i \geq 0.2] + \text{time}_{ij} \\ &\quad + I[\text{Age}_i < 50] + I[\text{Male}_i = 1], \text{ and} \\ \text{logit}(\text{missing}_{ij} | \text{Treatment}_i = \text{intervention}) &= -2.5 + I[\text{slope}_i < 0.5] + I[\text{Male}_i = 0]. \end{aligned}$$

One thousand simulated data sets were generated for each of nine scenarios.

Modeling approaches considered

A variety of methods were considered (Table 2). For reference, three models were fit: both mixed effects models ((1) the true or RI model for Scenarios A and B, and (2) the full RIRS model or true model for Scenario C that includes both random intercept and random slope) and (3) slope model (also the true model for Scenario C). In addition, the RI, RIRS, and slope models were fit to the incomplete data sets (for the RI and RIRS models, the incomplete data set consisted of any individual who contributed at least one outcome measurement and for the latter model, the incomplete data set included any individual who contributed at least two outcome measurements). The distinction between the slope and RIRS models lies in how the models were fit. In the former, slopes are derived and then linear regression is applied, regressing the derived slope on treatment arm and then using maximum likelihood methods to obtain estimates for the relevant parameter (treatment effect), whereas in the latter, a likelihood that incorporates normally distributed random effects for the slope and intercept is optimized to obtain estimates for the relevant parameter (a time by treatment interaction effect).

Five MI methods – 2 under the active approach, 2 under the passive approach, and 1 hybrid approach – were applied to the slope model. Under *Active imputation (Active)* the slope was derived and imputed as any other variable together with any missing underlying BMI measurements. In addition, slope is only imputed for those missing all 3 follow-up BMI measurements. Inconsistent values can then be obtained if the imputed slope does not match the imputed BMI measurements or when observed slope does not match all BMI measurements in cases where 1 or 2 BMI measurements are imputed. *Active imputation all (Active-*

MULTIPLE IMPUTATION FOR RATE OF CHANGE

All) is a variation on *Active* in which slope is only considered observed if all 4 BMI measurements contribute to its derivation. If slope is missing, it is imputed as any other variable as in *Active*. *Passive* and *Passive-All* involve deriving slope after imputing BMI measurements, where the variable slope can be considered for all variables other than those involved in its derivation. Thus, slope can be used in the imputation of covariates age and gender, for example, but it cannot be used in the imputation of the BMI measurements themselves.

After the imputation step, slope is derived from the underlying BMI measurements. In *Passive* this is done for those missing all 3 follow up measurements, and in *Passive-All* it is done for those missing at least one BMI measurement. As a consequence, internal inconsistencies may still be observed in *Passive* (but not in *Passive-All*) because slope is derived only for a proportion of subjects so that imputed BMI measurements may not match observed slope. The *Derive-Impute-Derive* approach is considered a hybrid approach of active and passive approaches. Slope is derived for those who contribute at least 2 measurements and imputed as any other variable as in *Active* and *Active-All*. After completing the imputation step, however, slope is re-derived from the underlying BMI measurements, some of which are observed and some of which are imputed. By including slope in the imputation model, relevant interrelationships among the variables are preserved. By re-deriving slope after the imputation process, internal consistency is gained. It is unclear, however, whether this step will interfere with what was gained in the imputation step by preserving interrelationships.

Ten imputations ($m=10$) were performed for each MI method for each data set. As passive approaches require imputation be performed using the FCS approach, FCS was used for all MI approaches. Procedures in Stata and R were used for this purpose.

Metrics for evaluation of methods

To evaluate performance of the methods, the relative mean squared error (rMSE) – or the mean squared error (MSE) relative to the MSE for the true model fitted to the full data set (gold standard) – was utilized, because this statistic considers both bias and efficiency. Other relevant statistics to gauge performance included the bias (the difference between the true (-0.5) and estimated parameter averaged over the number of simulations), the average model-based standard error, the empirical standard error, the square root of the variance of the coefficients across the simulations, the coverage probability (the percentage of time the true parameter is contained in the 95% confidence interval constructed using the

model-based standard error), and power (the percentage of time the null hypothesis that the parameter is zero is rejected using a two-sided Wald test at the 0.05 level of significance).

Table 2. Methods considered

Method	Description
RI Model Full	Mixed effects model with random intercept fit to the full data set
Slope Model Full	Slope model fit to the full data set
RIRS Model Full	Mixed effects model with random intercept and slope fit to the full data set
RI Model Incomplete	Mixed effects model with random intercept fit to those who provide at least one outcome measurement
Slope Model Incomplete	Slope model fit to those who provide at least two outcome measurements
RIRS Model Incomplete	Mixed effects model with random intercept and slope fit to those who provide at least one outcome measurement
MI Methods for Slope Model	
Active Imputation	Impute slope as any other variable for those who only have baseline outcome measurement
Active Imputation All	Impute slope as any other variable for those who do not have all four outcome measurements
Passive	Impute BMI using all variables except slope and derive slope for those with only baseline outcome measurements; slope is considered in imputation of variables other than those involved in derivation.
Passive All	Impute BMI using all variables except slope and derive slope for those missing any BMI measurements; slope is considered in imputation of variables other than those involved in derivation
Hybrid Derive-Impute-Derive	Derive slope and use in imputations of all variables and re-derive for consistency after imputation procedure

Results

The gold standard for comparisons is the true model fit to the full data set (RI model for Scenarios A and B and RIRS/slope model for Scenario C). The RIRS model is statistically equivalent to the slope model, although methods for fitting

MULTIPLE IMPUTATION FOR RATE OF CHANGE

differ – and thus, both the RIRS and slope models are considered the true model for Scenario C.

Performance of methods when data are generated from an RI model and data are MAR (Scenarios A1, B3, and B4)

Performance of RI, Slope, and RIRS models without imputation: When the data were MAR, both the RI model (true model) and the slope model were unbiased (true parameter = -0.5) when fit on the observed or incomplete data (e.g., average point estimates for Scenario A1 for all 3 models were -0.498 , See Table 3a). The slope model yielded least biased estimates (-0.004 versus -0.005 for the RI and RIRS models, Table 3a). Whereas the RI and RIRS models provided excellent coverage of 95.1% and 94.5 %, however, the slope model provided a lower coverage percentage of 85 (Figure 1). The RI and RIRS models had rMSEs of 6.5 and 6.6, whereas the slope model had an rMSE of 10.2.

Table 3a. Results from applying methods to Scenarios A1-A2

	Model/Method	Average Beta	Average SE	Bias	Cov	rMSE		
<i>RI, Slope, and RIRS Models Fit to Full Data from Scenario A (RI is true model; no auxiliary variables)</i>	RI (True)	-0.498	0.052	0.002	95.4	1.0		
	Slope	-0.498	0.052	0.002	95.4	1.0		
	RIRS	-0.498	0.052	0.002	95.9	1.0		
A1	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI (True)	-0.505	0.136	-0.005	95.1	6.5	
		Slope	-0.504	0.123	-0.004	85.4	10.2	
		RIRS	-0.505	0.136	-0.005	94.5	6.6	
	<i>Slope Models with MI</i>	Active	-0.460	0.130	0.040	87.1	9.9	
		Active-All	-0.443	0.583	0.057	99.6	22.2	
		Passive	-0.445	0.181	0.055	85.1	15.3	
		Passive-All	-0.391	0.275	0.109	77.5	41.5	
		Derive-Impute-Derive	-0.447	0.155	0.053	91.4	10.3	
	A2	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI (True)	-0.260	0.105	0.240	36.5	24.8
			Slope	-0.304	0.114	0.196	56.1	21.9
RIRS			-0.262	0.106	0.238	38.5	24.4	
<i>Slope Models with MI</i>		Active	-0.333	0.120	0.167	67.8	17.3	
		Active-All	-0.331	0.443	0.169	94.1	25.2	
		Passive	-0.243	0.118	0.257	41.9	28.8	
		Passive-All	-0.182	0.144	0.318	37.2	43.4	
		Derive-Impute-Derive	-0.331	0.121	0.169	66.3	17.5	

Figure 1. Coverage probability by method for Scenario A1, where data are generated from an RI model, data are MAR, and there are no auxiliary data

Slope model with MI approaches when no auxiliary information is available (Scenario A1):

When no auxiliary information was present and the data were MAR (Scenario A1) modest improvement in the slope model could be obtained through use of MI, but depended on the choice of MI method, where the rMSEs ranged from 9.9 (*Active*) to 41.5 (*Passive-All*) and where coverage ranged from 77.5% to 99.6% (Figure 1). Bias ranged from 0.04 (*Active*) to 0.11 (*Passive-All*). The “All” approaches suffered a loss in efficiency relative to their corresponding counterparts (e.g., *Active-All* gave an average standard error of 0.58 versus 0.13 for *Active*, and consequently yielded a high coverage of 99.6). With rMSEs of 9.9 and 10.3, the *Active* and *Derive-Impute-Derive* approaches were comparable to the slope model without MI, where *Active* MI provided slight improvement to the slope model without MI. Use of other MI methods with the slope model (*Active-All*, *Passive*, and *Passive-All*) gave worse results than using the slope model alone. The true (RI) and RIRS models without imputation performed best among all approaches, when considering bias, average SE, coverage and rMSE.

Slope model with MI approaches when auxiliary information is available (Scenarios B3 and B4):

Similar variability across MI methods was observed when auxiliary data were present (rMSEs varied from 8.7 to 20.5 in Scenario B3 and 6.8 to 26.2 in Scenario B4, Table 3b). There was a marked change, however. The inclusion of auxiliary terms in the imputation model greatly affected the performance of passive approaches. To see this, we compared results from

MULTIPLE IMPUTATION FOR RATE OF CHANGE

Table 3b. Results from applying methods to Scenarios B3-B7

		Model/ Method	Avg Beta	Avg SE	Bias	Cov	rMSE
RI, Slope, and RIRS Models Fit to Full Data from Scenario B (RI is true model; no auxiliary variables)		RI (True)	-0.498	0.052	0.002	94.8	1.0
		Slope	-0.498	0.052	0.002	94.8	1.0
		RIRS	-0.498	0.052	0.002	95.1	1.0
B3	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.503	0.136	-0.003	95.5	6.5
		Slope	-0.499	0.123	0.001	84.4	9.9
		RIRS	-0.503	0.136	-0.003	95.6	6.5
	Slope Models with MI	Active	-0.462	0.127	0.038	88.0	9.5
		Active-All	-0.440	0.588	0.060	100.0	20.5
		Passive	-0.456	0.134	0.044	88.8	8.7
		Passive-All	-0.428	0.183	0.072	80.0	20.1
		Derive-Impute-Derive	-0.446	0.154	0.054	90.8	10.3
B4	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.530	0.119	-0.030	94.5	5.2
		Slope	-0.481	0.122	0.019	87.9	8.5
		RIRS	-0.531	0.119	-0.031	94.0	5.2
	Slope Models with MI	Active	-0.492	0.127	0.008	90.6	8.1
		Active-All	-0.492	0.698	0.008	99.4	26.2
		Passive	-0.452	0.120	0.048	90.9	6.8
		Passive-All	-0.448	0.158	0.052	88.1	12.3
		Derive-Impute-Derive	-0.488	0.138	0.012	91.1	8.5
B5	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.259	0.105	0.241	36.0	24.2
		Slope	-0.310	0.114	0.190	58.2	20.7
		RIRS	-0.262	0.106	0.238	38.2	23.8
	Slope Models with MI	Active	-0.405	0.116	0.095	78.9	11.0
		Active-All	-0.403	0.503	0.097	96.6	20.4
		Passive	-0.368	0.099	0.132	72.1	10.0
		Passive-All	-0.331	0.105	0.169	60.6	14.7
		Derive-Impute-Derive	-0.402	0.117	0.098	79.4	11.3
B6	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.297	0.099	0.203	44.9	17.8
		Slope	-0.332	0.115	0.168	64.3	17.0
		RIRS	-0.301	0.100	0.199	47.9	17.2
	Slope Models with MI	Active	-0.426	0.117	0.074	86.0	9.0
		Active-All	-0.431	0.517	0.069	98.3	18.8
		Passive	-0.371	0.098	0.129	72.6	9.4
		Passive-All	-0.339	0.101	0.161	61.0	13.2
		Derive-Impute-Derive	-0.426	0.118	0.074	83.2	9.1
B7	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.390	0.064	0.110	59.1	5.8
		Slope	-0.485	0.083	0.015	93.0	2.8
		RIRS	-0.394	0.065	0.106	62.3	5.5
	Slope Models with MI	Active	-0.489	0.083	0.011	93.3	2.8
		Active-All	-0.486	0.498	0.014	100.0	10.2
		Passive	-0.487	0.080	0.013	93.2	2.6
		Passive-All	-0.417	0.063	0.083	73.3	3.9
		Derive-Impute-Derive	-0.490	0.083	0.010	92.6	2.8

Scenario A1 where there were no auxiliary variables to those from Scenarios B3 and B4 (Figure 2). In the latter scenarios, two auxiliary variables incorporated into the imputation process – one continuous and one dichotomous – were related to BMI measurements and the rate of change in BMI, and in Scenario B4 they were additionally related to missingness. *Passive* MI had a performance that was much improved (e.g., rMSE of *Passive* was 8.7 and 6.8 in Scenarios B3 and B4 versus 15.3 in Scenario A1; Figure 2, Table 3b) and provided estimates that were now competitive to those from *Active* MI; *Passive* MI yielded the lowest rMSEs of all MI methods and demonstrated improvement over the slope model without MI (rMSE of 8.7 versus 9.9 and 6.8 versus 8.5 for Scenarios B3 and B4, respectively).

The average standard error by bias across all methods for Scenario B4 is shown in Figure 3. A desirable method is one closest to the (0,0) point on the graph (and closest to the true (RI), RIRS, and slope models fitted to the full data) as this implies low bias and high efficiency. Although *Passive* in the presence of auxiliary terms is more competitive to *Active*, *Active* yields estimates with the least bias (e.g., in Scenario B4, bias for *Active* is 0.008 versus 0.048 for *Passive*). With rMSEs of 8.1 and 6.8, however, neither *Active* nor *Passive* MI applied to the slope model improved overall performance over the true (RI) model without imputation (rMSE=5.2).

Figure 2. Ratio of relative mean squared error for MI methods with auxiliary data applied to slope model to relative mean squared error for MI methods without auxiliary data applied to slope model

MULTIPLE IMPUTATION FOR RATE OF CHANGE

Figure 3. Bias by average standard error for Scenario B4, where data, generated from an RI model, are MAR, and auxiliary data are available

Performance of methods when data are generated from an RIRS/slope model and data are MAR (Scenario C8)

Performance of RI, Slope, and RIRS models without imputation: When the data were MAR, among models without imputation, the slope model produced the least biased results (-0.07 for the slope model and -0.097 and -0.092 for the RI and RIRS models; Table 3c). Coverage was lower for the RI model relative to the slope and RIRS models (84.3 versus 89.8 and 90.7 for the slope and RIRS models). The RIRS model gave the lowest rMSE of 3.4 with the RI and slope models providing comparable rMSEs of 3.8 and 4.2.

Slope model with MI approaches when auxiliary information is available (Scenario C8):

The rMSEs for the MI approaches applied to the slope model ranged from 3.0 (*Active*) to 8 (*Active-All*). *Active* MI yielded the least biased and most efficient estimates with *Passive* MI performing comparably. The results from fitting a slope model with either *Active* or *Passive* MI indicated improvement over the true models without imputation, where *Active* provided results with the least bias.

Table 3c. Results from applying methods to Scenarios C8-C9

		Model/Method	Average Beta	Average SE	Bias	Cov	rMSE
<i>RI, Slope, and RIRS Models Fit to Full Data from Scenario C (RI is true model; no auxiliary variables)</i>		RI	-0.510	0.069	-0.010	83.8	1.0
		Slope (True)	-0.510	0.096	-0.010	93.9	1.0
		RIRS (True)	-0.510	0.096	-0.010	93.8	1.0
C8	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI	-0.597	0.138	-0.097	84.3	3.8
		Slope (True)	-0.570	0.157	-0.070	89.8	4.2
		RIRS (True)	-0.592	0.154	-0.092	90.7	3.4
	<i>Slope Models with MI</i>	Active	-0.464	0.143	0.036	89.5	3.0
		Active-All	-0.450	0.693	0.050	99.7	8.0
		Passive	-0.456	0.152	0.044	91.7	3.2
		Passive-All	-0.453	0.196	0.047	90.7	6.0
		Derive-Impute- Derive	-0.461	0.153	0.039	91.4	3.1
	C9	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI	-0.159	0.111	0.341	20.4
Slope (True)			-0.253	0.150	0.247	60.7	9.5
RIRS (True)			-0.258	0.140	0.242	58.7	8.4
<i>Slope Models with MI</i>		Active	-0.444	0.149	0.056	89.4	3.2
		Active-All	-0.393	0.539	0.107	98.6	6.3
		Passive	-0.403	0.127	0.097	88.9	2.7
		Passive-All	-0.368	0.132	0.132	82.7	3.8
		Derive-Impute- Derive	-0.385	0.134	0.115	83.8	3.6

Performance of methods when data are generated from an RI model and data are NMAR (Scenarios A2, B5, and B6)

Performance of RI, Slope, and RIRS models without imputation: Under NMAR, the slope model yielded results with the least bias, highest coverage and lowest rMSE (Tables 3a and 3b). For example, in Scenario A2, bias ranged from 0.196 (slope model) to 0.24 (RI/true model), coverage ranged from 36.5 (RI/true model) to 56.1 (slope model), and rMSE ranged from 21.9 (slope model) to 24.8 (RI/true model).

Slope model with MI approaches when no auxiliary information is available (Scenario A2:)

As under MAR, there was considerable variability among MI methods applied to the slope model with rMSEs ranging from 17.3 to 43.4 (Table 3a). With no auxiliary information, passive approaches exhibited the most bias (0.257 and 0.318 for *Passive* and *Passive-All*) and the worst coverage probabilities (41.9 and 37.2). The active approaches (*Active* and *Active-All*) and

Derive-Impute-Derive proved superior to the passive approaches when considering both bias and efficiency.

Slope model with MI approaches when auxiliary information is available (Scenarios B5 and B6):

Improvement in MI methods from incorporating auxiliary data under NMAR was similar to that seen under MAR. Specifically, auxiliary terms greatly improved performances for the passive approaches and modestly for the *Active* and *Derive-Impute-Derive* approaches. All methods, however, had larger improvements when auxiliary variables were also related to missingness (Scenario B6). rMSEs ranged from 10 (*Passive*) to 20.4 (*Active-All*) for Scenario B5 and 9 (*Active*) to 18.8 (*Active-All*) for Scenario B6 and suggested that the performances of *Passive*, *Active*, and *Derive-Impute-Derive* were comparable overall. *Active* and *Derive-Impute-Derive*, however, achieved estimates with comparable and less bias than *Passive* (e.g., 0.095 and 0.098 versus 0.132 in Scenario B5) and higher coverage, where, for example, *Active* and *Derive-Impute-Derive* provided coverage estimates of 86% and 83.2% compared to *Passive*, which yielded coverage of 72.6% in Scenario B6. Figure 4 depicts efficiency and bias across MI methods and shows that under NMAR in the presence of auxiliary information, the true (RI) and slope models performed considerably worse with respect to bias and efficiency than the slope model with (and even without) MI. The *Active* and *Derive-Impute-Derive* approaches were the most preferable choices here (Figure 4).

Figure 4. Bias by average standard error for Scenario B6, where data, generated from an RI model, are NMAR, and auxiliary data are available

Performance of methods when data are generated from an RIRS model and data are NMAR (Scenario C9)

Performance of RI, Slope, and RIRS models without imputation: When the data were NMAR, rMSEs ranged from 8.4 (RIRS/true model) to 14.9 (RI model) (Table 3c). Coverage was low for the RI model (20.4%) and higher (although not at nominal levels) for the RIRS and slope models (58.7% and 60%). Bias ranged from 0.242 (RIRS model) to 0.341 (RI model).

Slope model with MI approaches when auxiliary information are available (Scenario C9):

When applying MI to the slope model and auxiliary information was present, rMSEs ranged from 2.7 (*Passive*) to 6.3 (*Active-All*) (Table 3c). Although *Passive* yielded the lowest rMSE, *Active* yielded the least biased estimate (0.056 compared to 0.097 for *Passive* and 0.115 for *Derive-Impute-Derive*) and the best coverage of the three competing methods (89.4%). Figure 5 plots the average standard error by the bias for each method. Points corresponding to the *Active*, *Derive-Impute-Derive*, and *Passive* methods were closer to those corresponding to the RI and true (RIRS and slope) models fitted to the full data set than the RI and true (RIRS and slope) models fitted to the incomplete data set, demonstrating the improvement of MI approaches that was not observed under MAR.

Figure 5. Bias by average standard error for Scenario C8, where data, generated from an RIRS model, are MAR, and auxiliary data are available

Impact of percentage missing under NMAR (comparison of Scenarios B6 and B7)

In Scenario B7, only 20% of the observations were missing on average. Similar patterns emerged in variability across approaches (Table 3b). The slope model without MI was superior to the true model (RI), with relative MSEs of 2.9 versus 5.8, achieved good coverage of 93% compared to only 59.1% for the RI (true) model and 62.3% for the RIRS model. Relative MSEs among MI methods applied to the slope model varied from 2.6 to 10.2. As in Scenario B6 where 40% of the observations were missing, improvement over the slope model without MI was gained with *Active*, *Derive-Impute-Derive*, and *Passive* methods. The advantage, however, was more modest than when the percentage missing was higher.

Conclusion

There was variability in results from applying various MI methods when slope was the outcome and missing for a proportion of subjects, and a two-stage linear regression was applied. When data were generated from an RI model, and the data were MAR, the RI model without imputation was the most appropriate choice with or without auxiliary data. Without auxiliary information, little was gained by doing MI, and in fact, considerable harm could come from using MI under a passive approach. When auxiliary information was present, the performance of *Passive* improved and surpassed that of *Derive-Impute-Derive* and *Active* in terms of the rMSE statistic; all things considered, however, *Active*, *Derive-Impute-Derive*, and *Passive* were comparable. All methods, however, were worse than the true RI model.

When the data were NMAR, much was gained by using a slope model over the RI and RIRS models, and importantly, MI with the slope model was considerably preferable over the slope model alone. When the data arose from either an RI or RIRS model, the slope model without MI outperformed the true model. Even without auxiliary data, *Active* and *Derive-Impute-Derive* with the slope model provided benefit (in terms of MSE and bias) over the models without MI, whereas the passive approaches proved harmful. With auxiliary information, however, *Passive* also provided improved estimates over the models without MI.

Although it is possible that the slope model is more robust to the violation of the MAR condition than the RI model, this difference in performance should be explored under various NMAR conditions, particularly in the absence of auxiliary data. In the presence of auxiliary information, it makes sense that, as we observed,

the slope model with MI outperformed all models fit to the incomplete data without MI because of the incorporation of auxiliary terms, making MAR a more reasonable assumption, and therefore a better choice under such conditions.

There were two *All* approaches considered, which were variations on the *Active* and *Passive* methods: *Active-All*, and *Passive-All*. Each performed worse than its counterpart. The idea behind the *All* approaches was that better slope measurements could be obtained by providing one based on a complete set of values (in this case, 4) rather than one based on only a partially observed trajectory, which may then yield less biased estimates of the relationship of interest.

A reduction in bias was not observed, however. Instead, the *All* approaches were more inefficient and more biased than their counterparts, where *Active-All* in particular suffered the largest loss in efficiency. The loss in efficiency may be explained by the application of these approaches to a modified data set with a higher proportion of missing data. More specifically, the *All* approaches assumed slope was missing for those missing at least 1 BMI measurement as opposed to for those missing all 3 follow-up BMI measurements, thereby increasing the fraction of missing information. Despite the increase in missing observations, the *All* method was applied with the same number of imputations ($m=10$) as with its counterpart method. Even when the number of imputations was doubled, however, these relative losses in efficiency remained (data not shown).

It may be that a more substantial increase in the number of imputations is needed, and this remains to be explored. In the active setting, such an objective simply adds a new variable to be imputed, while making use of the partially observed slope as an auxiliary term. While the objective of the passive approaches is to have internal consistency, full consistency was not achieved in *Passive*, because the observed slope and not the derived slope was used for subjects with at least 2 BMI outcome measurements. Although full internal consistency was achieved in *Passive-All* gains in properties of the estimates were not observed. These approaches likely failed because the focus was incorrectly on producing imputed values closer to what would have been observed (i.e., by making them consistent) with the idea that such values would lead to better descriptions of relationships rather than on exploiting the interrelationships among the variables. It was demonstrated that use of observed slope (even if based on only a fraction of measurements) produced estimates of relevant parameters with superior properties to those based on a full number of measurements where a fraction of those measurements were imputed. Findings may vary, however, in settings with a larger number of longitudinal measurements.

MULTIPLE IMPUTATION FOR RATE OF CHANGE

Derive-Impute-Derive behaved comparably to *Active* MI in all scenarios including in the presence or absence of auxiliary information. An active-type approach that requires an extra step of editing or re-deriving for consistency, *Derive-Impute-Derive* did not demonstrate meaningful improvements over *Active*. Larger variation in performance may be observed in settings with higher proportions of internal consistency. This would need to be explored in further simulation studies.

Variability between *Active/Derive-Impute-Derive* MI and passive approaches was largely due to the presence of auxiliary information. Although some improvement was observed with *Active* and *Derive-Impute-Derive* with the presence of auxiliary information, it was modest relative to the improvement observed among the passive approaches. It was the auxiliary information that allowed *Passive* to be competitive to *Active* and *Derive-Impute-Derive*; because observed slope is not considered an auxiliary term when passively imputing BMI, other auxiliary information is critical for gleaning the interrelationships among relevant variables in the model. In contrast, in *Active* and *Derive-Impute-Derive*, BMI measurements (and other variables) are used as auxiliary terms when imputing slope and vice versa and the interrelationships among these terms are jointly considered. It would be interesting to further explore the differences in performance of approaches in settings where other covariates have missing values.

There were some computational issues to consider when implementing the passive and hybrid approaches. Unlike *Active*, they required some extra steps beyond the usual MI programming. *Derive-Impute-Derive* requires re-deriving slope to ensure internal consistency after doing an active-type imputation (See Appendix A for STATA code). *Passive* – possible in STATA and R, but not in SAS – requires an explicit statement to define the derived term as a function of the main effects so that updating of the derived variable can be performed during the imputation procedure (See Appendix B). With a derived term that requires more than a basic operation of addition or multiplication as in interaction effects, this can be challenging. If the derivation requires more than one line, this may be problematic in STATA. Thus, if performances across MI methods are equal, *Active* is preferred for its computational simplicity. Further, it is important to note that *Active* – unlike *Passive* – is available in all MI packages including R, STATA, SAS and SPSS.

It may be argued that the imputation methods considered here do not lead to proper imputations. Rubin provides criteria required (i.e., properties necessary for the imputation method) to yield imputations that are deemed proper (Rubin, 1987). Consequently, estimates of population quantities that incorporate proper multiple

imputations are statistically valid (unbiased with appropriate coverage). One such assumption requires congeniality of the imputation and scientific models, where congeniality implies that the imputed data can be derived from the scientific model (Meng, 1994).

An example of an un-congenial model may be one where the scientific model is a linear regression of BMI on treatment arm and age, where BMI is missing for a proportion of subjects, and the imputation model assumes multivariate normality of BMI, treatment arm, and possibly other auxiliary terms, but excludes age. Because the scientific model considers interrelationships between age and BMI and treatment arm, and the imputation model does not, the models are said to be un-congenial. In this case, when the slope model is the scientific model, recall that there are two stages involved: the derivation of the slopes (Equation 1a) followed by the regression of the derived slopes on treatment arm (Equation 1b).

Because the imputation model does not incorporate uncertainty of the parameters in Equation 1a, it may be considered un-congenial. Imputing under a fully Bayesian paradigm that puts priors on all the parameters listed in both Equations 1a and 1b when deriving the posterior predictive distribution would be considered proper. MI methods considered here, however, were only those that could be implemented using mainstream software like Stata, R, and SAS, with minimal additional code, and specifically where the slope model was the scientific model of interest. Our reference for ideal performance was the true model fit to the data set with no missing values, rather than the ideal MI method.

Further, methods that incorporate MI with mixed effects models such as PAN, developed by Schafer (Schafer, 1997a; Schafer, 1997b), were not considered. Nor was a shared parameter model considered, which is also appropriate for longitudinal models and allows the missing data mechanism to be NMAR (e.g., Wu & Bailey, 1989). The former approach can be implemented using a package in R, whereas code for fitting the shared parameter model is not available in mainstream software packages. It would be interesting to compare MI methods under both the slope model and mixed effects models in the future. The current focus, however, was specifically on fitting the slope model, where the mixed effects models fit to the full and incomplete data without MI served as references. Clinical investigators are often averse to applying mixed effects models if they can apply a seemingly simpler two-stage linear regression model. This is the case with Stanford GOALS. Thus, the intent was to evaluate practical MI methods under the slope model.

MULTIPLE IMPUTATION FOR RATE OF CHANGE

The goal was to provide a set of guidelines on how to analyze rate of change in the presence of missing data. [Figure 6](#) provides a flow chart to aid in decision-making. Typically, the first step when faced with missing data is to consider plausible missing data mechanisms.

We take one more step back, however, and ask to consider from which model the data may be reasonably generated. In this study, data were generated from both an RI model and the RIRS model. The former assumes each individual has a BMI around which a series of BMI values center. Further it assumes that how those BMI values change over time may be affected by treatment regimen. Data arising from an RIRS model implies each individual additionally has his/her own rate of change in BMI, and, as before, trajectories may be affected by a treatment regimen. It leads to slight differences in recommendations. If the data are likely to arise from a slope model (i.e., if you can assume BMI values arise from a rate of change), we recommend using a slope model with *Active* MI (if either MAR or NMAR, and in the presence of auxiliary information).

Although some improvements were observed with *Passive* over *Active* when auxiliary data were present, *Active* was still less biased, comparably efficient and simpler to implement. If auxiliary information are not available, a more sophisticated approach should be considered such as the shared parameter model as nominal coverage was not achieved with the methods considered here ([Duch & Stevenson, 2005](#)). If the data arise from an RI model, and if the data are MAR, the best choice among those considered is an RI model. Thus, investigators should reconsider specifying the scientific model as the slope model in favor of the RI model; the slope model without MI produced estimates with inferior properties. Further, in the absence of auxiliary information, little was gained with using MI. Some investigators, however, may be wary to fit and interpret parameters from a mixed effects model and may wish for the simplicity of a linear regression model (e.g., [Liao et al., 2013](#)) the benefits of which are described here ([Rogosa & Saner, 1995](#); [Gelman, 2005](#)). In this case, *Active* MI can improve upon the slope model without MI in the presence of auxiliary information. If the data are NMAR, the slope model with *Active* MI provides improvement over the mixed effects models in the presence of auxiliary terms. In general, however, if the data are NMAR, a more sophisticated approach such as the shared parameter model should be considered.

Figure 6. Flow chart of model choices

In longitudinal clinical trial settings such as Stanford GOALS it is likely the case that the data are NMAR. For example, those who respond well to the behavioral intervention may be more likely to have outcome values and intermittent BMI values. The availability of numerous solid auxiliary variables, however, is also likely, and conditional on these, it may be reasonable to assume the data are MAR. Baseline measurements of BMI, other adiposity measures, lipid profiles, as well as intermittent values of BMI are potentially useful auxiliary terms that can be incorporated into the analysis. Thus, if NMAR, standard MI methods that incorporate such terms can provide benefit over the mixed effects and slope models without MI. Although it is important to pre-specify a primary method of analyses (in Stanford GOALS, the primary approach is the slope model with *Active* MI), it is recommended that sensitivity analyses be performed – in particular, the mixed effects and slope models without MI, and the slope model with *Derive-Impute-Derive*, and *Passive* approaches. If results are comparable, this bolsters the findings from the primary analysis. If results are heterogeneous, it is critical to report the discrepancy in order to provide a context for the audience in which to interpret the findings.

Acknowledgements

The work described was supported in part by the Child Health Research Institute at Stanford University, by Award Number U01HL103629 from the National Heart, Lung, and Blood Institute, and by the Patient-Centered Outcomes Research Institute Grant ME-1303-5989. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the position or policy of the Department of Veterans Affairs, Patient-Centered Outcomes Research Institute, or the United States government.

References

- Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.
- Collins, L. M., Schafer, J. L., & Kam, C.M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6(4), 330-51. doi: [10.1037/1082-989X.6.4.330](https://doi.org/10.1037/1082-989X.6.4.330)
- Duch, R. M. and Stevenson, R. (2005). Context and the economic vote: a multilevel analysis. *Political Analysis*, 13(4), 387-409. doi: [10.1093/pan/mpi028](https://doi.org/10.1093/pan/mpi028)
- Gelman, A. (2005). Two-stage regression and multilevel modeling: a commentary. *Political Analysis*, 13(4), 459-461. doi: [10.1093/pan/mpi032](https://doi.org/10.1093/pan/mpi032)
- Jusko, K. L. and Shively, W. P. (2013). Applying a two-step strategy to the analysis of cross-national public opinion data. *Political Analysis*, 13(4), 327-344. doi: [10.1093/pan/mpi030](https://doi.org/10.1093/pan/mpi030)
- Liao, J., Brunner, E. J., & Kumari, M. (2013). Is there an association between work stress and diurnal cortisol patterns? Findings from the Whitehall II study. *PLoS One*, 8(12). doi: [10.1371/journal.pone.0081020](https://doi.org/10.1371/journal.pone.0081020)
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd ed. New York: J. Wiley & Sons.
- Meng, X-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9, 538-558.
- Mitani, A., Kurian, A., Das, A., & Desai, M. (2015). Navigating choices when applying multiple imputation in the presence of multi-level categorical interaction effects. *Statistical Methodology*, 27, 82-99. doi: [10.1016/j.stamet.2015.06.001](https://doi.org/10.1016/j.stamet.2015.06.001)

Robinson, T. N., Matheson, D., Desai, M., Wilson, D. M., Weintraub, D. L., Haskell, W. L., McClain, A., McClure, S., Banda, J. A., Sanders, L. M., Haydel, K. F., & Killen, J. D. (2013). Family, community and clinic collaboration to treat overweight and obese children: Stanford GOALS – A randomized controlled trial of a three-year, multi-component, multi-level, multi-setting intervention.

Contemporary Clinical Trials, 36(2), 421-435. doi: [10.1016/j.cct.2013.09.001](https://doi.org/10.1016/j.cct.2013.09.001)

Rogosa, D. & Saner, H. (1995). Longitudinal data analysis examples with random coefficient models. *Journal of Educational and Behavioral Statistics*, 20(2), 149-170. doi: [10.3102/10769986020002149](https://doi.org/10.3102/10769986020002149)

Royston, P. (2009). Multiple imputation of missing values: further update of ice, with an emphasis on categorical variables. *Stata Journal*, 9(3), 466-77.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-89. doi: [10.2307/2291635](https://doi.org/10.2307/2291635)

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons.

Schafer, J. L. (1997a). *Imputation of missing covariates under a general linear mixed model* (Technical Report). State College, PA: Dept. of Statistics, Penn State University.

Schafer, J. L. (1997b). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Seaman, S. R., Bartlett, J. W., White, I. R. (2012). Multiple imputation of missing covariates with non-linear effects and interactions: an evaluation of statistical methods. *BMC Medical Research Methodology*, 12:46. doi: [10.1186/1471-2288-12-46](https://doi.org/10.1186/1471-2288-12-46)

von Hippel, P. T. (2009). How to impute interactions, squares and other transformed variables. *Sociological Methodology*, 39, 265-91.

van Buuren, S. (2012). *Flexible imputation of missing data*. London: Chapman & Hall/CRC.

van Buuren S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, 16(3), 219-42. doi: [10.1177/0962280206074463](https://doi.org/10.1177/0962280206074463)

White, I. R., Royston, P., Wood, A. M. (2011). Multiple imputation using chained equations: issues and guidance for practice. *Statistics in Medicine*, 30(4), 377-99. doi: [10.1002/sim.4067](https://doi.org/10.1002/sim.4067)

MULTIPLE IMPUTATION FOR RATE OF CHANGE

Wu, M. C. & Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 45, 939-955.

Appendix A: *Derive-Impute-Derive* Imputation of Slope Using ICE

```

* Read in ith data set
insheet using "/Users/mdesai/Documents/COPTR/Multiple Imputation
Research/Simulated Data Sets/reducedwidescenA`i'.csv", clear

destring, ignore("NA") replace

* Derive and Impute:
* Doing Active imputation by including slope in the imputation model as any
variable

ice reducedslope bmi0 bmi1 bmi2 bmi3 txgroup age male,
saving("/Users/mdesai/Documents/nimpute.dta") m(10) replace

use "/Users/mdesai/Documents/nimpute.dta", clear

* Reshape data set from wide format to long format so that slope can be re-
derived
reshape long bmi, i(id _mj) j(visit)

* Derive again
statsby _b, by(id _mj) saving("/Users/mdesai/Documents/newb.dta", replace): reg
bmi visit

* Reshape from long to wide format and sort for analysis
reshape wide bmi, i(id _mj) j(visit)

use "/Users/mdesai/Documents/newb.dta", clear
sort id _mj

use "/Users/mdesai/Documents/nimpute.dta", clear
sort id _mj

merge 1:1 id _mj using "/Users/mdesai/Documents/newb.dta"

save "/Users/mdesai/Documents/merge.dta", replace

use "/Users/mdesai/Documents/merge.dta", clear

sort _mj id

* Now fit the scientific model using the re-derived outcome and summarize using
MICOMBINE
* where the newly derived term is called _b_visit

micombine reg _b_visit txgroup

```


Appendix B: *Passive* and *Passive-All* Imputation of Slope Using ICE

```
* read in the ith data set

insheet using "/Users/mdesai/Documents/COPTR/Multiple Imputation
Research/Simulated Data Sets/reducedwidescenA`i'.csv", clear

destring, ignore("NA") replace

* copy reducedslope
gen newslope = reducedslope

* passively impute reducedslope using ICE

ice bmi0 bmi1 bmi2 bmi3 txgroup age male newslope, passive(newslope: (-
1.5*(bmi0-(bmi0+bmi1+bmi2+bmi3)/4) + -0.5*(bmi1-(bmi0+bmi1+bmi2+bmi3)/4) +
0.5*(bmi2-(bmi0+bmi1+bmi2+bmi3)/4) + 1.5*(bmi3-(bmi0+bmi1+bmi2+bmi3)/4))/5)
saving(/Users/mdesai/Documents/nimpute.dta) m(10) seed(1234) replace
use "/Users/mdesai/Documents/nimpute.dta", clear

* Fit scientific model for Passive-All and summarize findings using MICOMBINE.
* Since newslope is used, it replaces the slope based on partially observed BMI
values

micombine reg newslope txgroup

*Fit scientific model for Passive and summarize findings using MICOMBINE.
*Here the derived slope is only used when all 3 BMI measurements are missing.

use "/Users/mdesai/Documents/nimpute.dta",clear
replace reducedslope = newslope if missing(reducedslope)
micombine reg reducedslope txgroup
```

Appendix C: Full Results from Applying Methods to Scenarios A1-C9 (Table A1)

		Model/ Method	Avg Beta	Avg SE	Emp SE	Avg Bias	Cov	MSE	rMSE	Power	
RI, Slope, and RIRS Models Fit to Full Data from Scenario A (RI is true model; no auxiliary variables)			RI (True)	-0.498	0.052	0.053	0.002	95.4	2.8	1.0	100.0
			Slope	-0.498	0.052	0.053	0.002	95.4	2.8	1.0	100.0
			RIRS	-0.498	0.052	0.053	0.002	95.9	2.8	1.0	100.0
A1	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.505	0.136	0.134	-0.005	95.1	18.0	6.5	95.9	
		Slope	-0.504	0.123	0.167	-0.004	85.4	28.0	10.2	93.9	
		RIRS	-0.505	0.136	0.135	-0.005	94.5	18.1	6.6	96.1	
	Slope Models with MI	Active	-0.460	0.130	0.160	0.040	87.1	27.3	9.9	89.0	
		Active-All	-0.443	0.583	0.241	0.057	99.6	61.3	22.2	1.3	
		Passive	-0.445	0.181	0.198	0.055	85.1	42.2	15.3	72.2	
		Passive-All	-0.391	0.275	0.320	0.109	77.5	114.4	41.5	34.4	
		Derive-Impute-Derive	-0.447	0.155	0.160	0.053	91.4	28.3	10.3	80.7	
		A2	RI (True)	-0.260	0.105	0.104	0.240	36.5	68.5	24.8	69.2
			Slope	-0.304	0.114	0.148	0.196	56.1	60.3	21.9	69.3
RIRS	-0.262		0.106	0.104	0.238	38.5	67.2	24.4	69.9		
Active	-0.333		0.120	0.141	0.167	67.8	47.7	17.3	74.1		
Slope Models with MI	Active-All	-0.331	0.443	0.202	0.169	94.1	69.6	25.2	0.5		
	Passive	-0.243	0.118	0.115	0.257	41.9	79.5	28.8	54.7		
	Passive-All	-0.182	0.144	0.136	0.318	37.2	119.6	43.4	26.1		
	Derive-Impute-Derive	-0.331	0.121	0.140	0.169	66.3	48.2	17.5	74.3		
RI, Slope, and RIRS Models Fit to Full Data from Scenario B (RI is true model; no auxiliary variables)			RI (True)	-0.498	0.052	0.053	0.002	94.8	2.8	1.0	100.0
			Slope	-0.498	0.052	0.053	0.002	94.8	2.8	1.0	100.0
			RIRS	-0.498	0.052	0.053	0.002	95.1	2.8	1.0	100.0
B3	RI, Slope, and RIRS Models Fit to Incomplete Data	RI (True)	-0.503	0.136	0.136	-0.003	95.5	18.4	6.5	95.1	
		Slope	-0.499	0.123	0.168	0.001	84.4	28.2	9.9	92.9	
		RIRS	-0.503	0.136	0.136	-0.003	95.6	18.4	6.5	95.2	
	Slope Models with MI	Active	-0.462	0.127	0.157	0.038	88.0	26.1	9.5	91.3	
		Active-All	-0.440	0.588	0.230	0.060	100.0	56.5	20.5	0.7	
		Passive	-0.456	0.134	0.150	0.044	88.8	24.4	8.7	90.1	
		Passive-All	-0.428	0.183	0.226	0.072	80.0	56.4	20.1	66.4	
		Derive-Impute-Derive	-0.446	0.154	0.159	0.054	90.8	28.3	10.3	81.3	
		B4	RI (True)	-0.530	0.119	0.118	-0.030	94.5	14.7	5.2	99.2
			Slope	-0.481	0.122	0.155	0.019	87.9	24.2	8.5	94.3
RIRS	-0.531		0.119	0.118	-0.031	94.0	14.8	5.2	99.2		
Active	-0.492		0.127	0.151	0.008	90.6	22.7	8.1	95.0		
Slope Models with MI	Active-All	-0.492	0.698	0.271	0.008	99.4	73.3	26.2	0.9		
	Passive	-0.452	0.120	0.130	0.048	90.9	19.1	6.8	94.4		
	Passive-All	-0.448	0.158	0.178	0.052	88.1	34.5	12.3	79.0		
	Derive-Impute-Derive	-0.488	0.138	0.154	0.012	91.1	23.9	8.5	92.9		

MULTIPLE IMPUTATION FOR RATE OF CHANGE

Table A1 continued

		Model/ Method	Avg Beta	Avg SE	Emp SE	Avg Bias	Cov	MSE	rMSE	Power
B5	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI (True)	-0.259	0.105	0.105	0.241	36.0	68.9	24.2	70.0
		Slope	-0.310	0.114	0.152	0.190	58.2	58.9	20.7	72.5
		RIRS	-0.262	0.106	0.106	0.238	38.2	67.6	23.8	70.2
	<i>Slope Models with MI</i>	Active	-0.405	0.116	0.147	0.095	78.9	30.4	11.0	88.5
		Active-All	-0.403	0.503	0.218	0.097	96.6	57.0	20.4	0.4
		Passive	-0.368	0.099	0.103	0.132	72.1	28.0	10.0	94.7
		Passive-All	-0.331	0.105	0.112	0.169	60.6	41.2	14.7	87.1
		Derive-Impute-Derive	-0.402	0.117	0.148	0.098	79.4	31.6	11.3	86.5
B6	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI (True)	-0.297	0.099	0.096	0.203	44.9	50.6	17.8	86.5
		Slope	-0.332	0.115	0.143	0.168	64.3	48.4	17.0	76.1
		RIRS	-0.301	0.100	0.098	0.199	47.9	48.9	17.2	87.1
	<i>Slope Models with MI</i>	Active	-0.426	0.117	0.140	0.074	86.0	25.2	9.0	92.7
		Active-All	-0.431	0.517	0.219	0.069	98.3	52.7	18.8	1.2
		Passive	-0.371	0.098	0.099	0.129	72.6	26.4	9.4	96.9
		Passive-All	-0.339	0.101	0.105	0.161	61.0	37.0	13.2	89.8
		Derive-Impute-Derive	-0.426	0.118	0.142	0.074	83.2	25.6	9.1	91.5
B7	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI (True)	-0.390	0.064	0.066	0.110	59.1	16.4	5.8	100.0
		Slope	-0.485	0.083	0.089	0.015	93.0	8.1	2.8	99.9
		RIRS	-0.394	0.065	0.066	0.106	62.3	15.6	5.5	100.0
	<i>Slope Models with MI</i>	Active	-0.489	0.083	0.088	0.011	93.3	7.9	2.8	99.9
		Active-All	-0.486	0.498	0.168	0.014	100.0	28.4	10.2	2.4
		Passive	-0.487	0.080	0.085	0.013	93.2	7.4	2.6	99.9
		Passive-All	-0.417	0.063	0.064	0.083	73.3	11.0	3.9	100.0
		Derive-Impute-Derive	-0.490	0.083	0.089	0.010	92.6	7.9	2.8	99.9
C8	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI	-0.510	0.069	0.096	-0.010	83.8	9.3	1.0	100.0
		Slope (True)	-0.510	0.096	0.096	-0.010	93.9	9.3	1.0	100.0
		RIRS (True)	-0.510	0.096	0.096	-0.010	93.8	9.3	1.0	100.0
	<i>Slope Models with MI</i>	Active	-0.597	0.138	0.161	-0.097	84.3	35.3	3.8	97.4
		Slope (True)	-0.570	0.157	0.185	-0.070	89.8	39.2	4.2	90.7
		RIRS (True)	-0.592	0.154	0.153	-0.092	90.7	31.8	3.4	96.9
		Active-All	-0.464	0.143	0.164	0.036	89.5	28.1	3.0	85.9
		Passive	-0.456	0.152	0.166	0.044	91.7	29.5	3.2	83.8
C9	<i>RI, Slope, and RIRS Models Fit to Incomplete Data</i>	RI	-0.159	0.111	0.147	0.341	20.4	137.9	14.9	36.5
		Slope (True)	-0.253	0.150	0.164	0.247	60.7	87.9	9.5	40.6
		RIRS (True)	-0.258	0.140	0.138	0.242	58.7	77.5	8.4	46.8
	<i>Slope Models with MI</i>	Active	-0.444	0.149	0.164	0.056	89.4	30.0	3.2	81.2
		Active-All	-0.393	0.539	0.217	0.107	98.6	58.4	6.3	0.3
		Passive	-0.403	0.127	0.127	0.097	88.9	25.5	2.7	89.4
		Passive-All	-0.368	0.132	0.133	0.132	82.7	35.0	3.8	79.0
		Derive-Impute-Derive	-0.385	0.134	0.143	0.115	83.8	33.7	3.6	80.6

Construction of Pair wise Balanced Design

**Rajarathinam
Arunachalam**

Manonmaniam Sundaranar
University
Tirunelveli, India

**Mahalakshmi
Sivasubramanian**

Manonmaniam Sundaranar
University
Tirunelveli, India

Dilip Kumar Ghosh
Saurashtra University
Gujarat, India

A new procedure for construction of a pair wise balanced design with equal replication and un-equal block sizes based on factorial design are presented. Numerical illustration also provided. It was found that the constructed pair wise balanced design was found to be universal optimal.

Keywords: C-matrix, Balanced Design, Connected, Equal replicated and unequal block sizes, Eigenvalues, Factorial Design.

Introduction

Among all the incomplete block design Balanced Incomplete Block Design (BIBD) is the simply easiest and most suitable incomplete block design in regards of construction and analysis. Because BIBD is binary, connected, proper, equi-replicated, balanced and non orthogonal. However, BIBD is not available for all parameters, because BIBD exist only if

- a) $vr = bk$
- b) $\lambda(v - 1) = r(k - 1)$ and
- c) $b \geq v$

holds true. So in place of BIBD, we need another incomplete block design which is balanced. This type of incomplete block design is called variance balanced; efficiency balanced and pair wise balanced designs.

In a pair wise balanced block design, any pair of treatments within the blocks occur equally often. The literature on combinatorial theory contains many

---

Dr. A. Rajarathinam is a Professor of Statistics. Email him at: arrathinam@yahoo.com. Email Mahalakshmi Sivasubramanian email at: seethamaha06@gmail.com. Dilip Kumar Ghosh is in the Department of Statistics. Email at: ghosh_dkg@yahoo.com.

PAIR WISE BALANCED DESIGN

contributions to the existence and construction of pair wise balanced designs, including for settings with unequal block sizes.

Bose and Shrikhande (1959a) discussed about the pair wise balanced design in addition to existence of orthogonal Latin square designs. Bose and Shrikhande (1960a) obtained the various methods for the construction of pair wise orthogonal sets of Latin square design. However, the detailed discussion and construction on pair wise balanced design is studied by Bose and Shrikhande (1960a). In other words, we can say that Bose and Shrikhande (1959a, 1959b, 1960a) introduced a general class of incomplete block design which they called pair wise balanced design of index λ . The pair wise balanced design also share some of the properties of BIB designs like every pair of treatments occurs together in λ blocks. The concept of pair wise balanced design is merely the combinatorial interest in block designs. Because with the help of pair wise balanced design many other incomplete block designs can be constructed. For example, Bose and Shrikhande (1960b) used the pair wise balanced design in the context of constructing Mutually Orthogonal Latin Square (MOLS). Hedayat and Stufken (1989) showed that the problems of constructing pair wise balanced designs and variance balanced block designs are equivalent. Effanga, Ugboh, Enang and Eno (2009) developed a non-linear non-preemptive binary integer goal programming model for the construction of D -optimal pair wise balanced incomplete block designs. The literature on combinatorial theory contains many contributions to the existence and construction of pair wise balanced designs with un-equal block sizes.

Bose and Shrikhande (1960b) defined pair wise balanced design as following:

Definition: An arrangement of v treatments in b blocks is defined as pair wise balanced design of index λ of type $(v; k_1, k_2, \dots, k_m)$ provided

- i) Each set contains $(k_1, k_2, \dots, k_m)$ symbols that are all distinct
- ii) $k_i \leq v; k_i \neq k_j$ and
- iii) every pair of distinct treatments occurs in exactly λ sets of the design.

Further they also showed some of the parametric relation of pair wise balanced design which is namely:

- i) $b = \sum_{i=1}^m b_i$
- ii) $\lambda v(v-1) = \sum_{i=1}^m b_i k_i (k_i - 1)$

A characterization of pair wise balanced design in terms of NN^T matrix can be expressed in the following way.

A block design D is called pair wise balanced design if all the off diagonal elements of NN^T matrix are same (constant) i.e.,

$$NN^T = (r - \lambda)I_v + \lambda E_{vv} \quad (1)$$

where I_v is an identity matrix of order v and E_{vv} is the unit matrix of order $(v \times v)$.

Methods of Construction

Pair wise balanced design using 2^n symmetrical factorial design deleting the control treatment and merging all n main effects

Here we discuss the construction of unequal block sizes and equi-replicated binary pair wise balanced design from symmetrical factorial designs. First of all we will write a lemma without proof.

Lemma. Let us consider a 2^n symmetrical factorial experiment.

- (i) from these 2^n treatment combinations let delete the control treatments.
- (ii) merge all those treatment combinations which represent n main effects and further consider these n merged treatment combinations as one treatment combinations. This way, we have $2^n - n$ treatment combinations
- (iii) finally consider $(2^n - n)$ treatment combinations as the blocks for the required design.

We prove this lemma using following example. Let $n = 3$. The $2^3 = 8$ treatment combinations are

PAIR WISE BALANCED DESIGN

$$\begin{array}{ccc}
 0 & 0 & 0 \\
 1 & 0 & 0 \\
 0 & 1 & 0 \\
 1 & 1 & 0 \\
 0 & 0 & 1 \\
 1 & 0 & 1 \\
 0 & 1 & 1 \\
 1 & 1 & 1
 \end{array}$$

Delete the control treatment. Next merge all treatment combinations where level of one factor is one while level of other factor is zero. Keep the remaining treatment combinations as such. Finally, the treatment combinations are

$$\begin{array}{ccc}
 1 & 1 & 1 \\
 1 & 1 & 0 \\
 1 & 0 & 1 \\
 0 & 1 & 1 \\
 1 & 1 & 1
 \end{array}$$

Call this matrix as M which is the combinations of $2^n - n$ treatment combinations. The method of constructing pair wise balanced design is discussed in the following theorem.

Theorem. If there exist 2^n symmetrical factorial experiments then there always exist unequal block sizes, equi-replicated, binary pair wise balanced design, by deleting the control treatment and merging all the treatment combinations belonging to main effects and then considering these merged n treatment combinations as one, with the following parameters

$$v = n, b = 2^n - n, r = 2^{n-1}, \underline{k} = \left(\underbrace{2, \dots, 2}_{nC_2}; \underbrace{3, \dots, 3}_{nC_3}; \underbrace{n-1, \dots, n-1}_{nC_{n-1}}; n, n \right)' \text{ and } \lambda = 2^{n-2} + 1$$

Proof: Let us consider a 2^n symmetrical factorial experiment. This has 2^n treatment combinations. Consider n factors as rows and 2^n treatment combinations as columns. Now using the above lemma, we have the following incidence matrix N of a design d .

$$N = \begin{bmatrix} 1 & 1 & 1 & 0 & \cdots & \cdots & 0 & 1 \\ 1 & 1 & 0 & 1 & \cdots & \cdots & 1 & 1 \\ 1 & 0 & 1 & 1 & \cdots & \cdots & 1 & 1 \\ 1 & 0 & 0 & 0 & \cdots & \cdots & 1 & 1 \\ 1 & 0 & 0 & 0 & \cdots & \cdots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 1 & 0 & 0 & 0 & \cdots & \cdots & 1 & 1 \\ 1 & 0 & 0 & 0 & \cdots & \cdots & 1 & 1 \end{bmatrix} \quad (2)$$

Since we have n rows and considered these as treatments, obviously $v = n$.

For incidence matrix N , among ${}^nC_2$ columns, in each column element '1' will occur two times and 0 will occur $(n - 2)$ times. Similarly for ${}^nC_3$ columns, element '1' will occur three times and element 0 will occur $(n - 3)$ times in each column, and so on, i.e., for ${}^nC_{n-1}$ columns, element '1' will occur $(n - 1)$ time and element 0 will occur 1 time. Moreover there will be one column whose all elements are unity. Finally one more column, which is obtained by merging n treatment combinations, is having also all elements are '1'. Hence numbers of blocks are

$$\begin{aligned} b &= {}^nC_2 + {}^nC_3 + \dots + {}^nC_{n-1} + 1 + 1 \\ &= {}^nC_0 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_{n-1} + {}^nC_n \\ &= {}^nC_0 + {}^nC_1 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n - {}^nC_1 \end{aligned}$$

Since ${}^nC_0 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n = 2^n$ and therefore $b = 2^n - {}^nC_1 = 2^n - n$.

Again each row of N contains 2^{n-1} times '1' and $(2^{n-1} - n)$ times "zero" and hence $r = 2^{n-1}$. Obviously block size is

PAIR WISE BALANCED DESIGN

$$\underline{k} = \left(\underbrace{2, \dots, 2}_{nC_2}; \underbrace{3, \dots, 3}_{nC_3}; \underbrace{n-1, \dots, n-1}_{nC_{n-1}}; n, n \right)'$$

Finally the number of treatments $v = n$, number of blocks $b = 2^n - n$ and the number of replication is $r = 2^{n-1}$.

Using the incidence matrix N shown in (2), we have the following C -matrix.

$$C = \begin{pmatrix} \alpha & \beta & \beta & \beta & \cdots & \beta \\ \beta & \alpha & \beta & \beta & \cdots & \beta \\ \beta & \beta & \alpha & \beta & \cdots & \beta \\ \beta & \beta & \beta & \alpha & \cdots & \beta \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \beta & \beta & \cdots & \alpha \end{pmatrix}_{(v \times v)} \quad (3)$$

where $\alpha = r - \sum_j \frac{n_{ij}^2}{k_j}$ and $\beta = -\sum_j \frac{n_{ij}n_{ij}}{k_j}$

The eigenvalues of the above C -matrix are obtained as

$$\theta = \alpha \left[\frac{v}{v-1} \right]$$

with multiplicities $(v-1)$. For the design having the incidence matrix N given in equation (2) the NN^T is given below.

$$NN^T = \begin{bmatrix} 2^{n-1} & 2^{n-2} + 1 & \cdots & 2^{n-2} + 1 \\ 2^{n-2} + 1 & 2^{n-1} & \cdots & 2^{n-2} + 1 \\ \vdots & \vdots & \ddots & \vdots \\ 2^{n-2} + 1 & 2^{n-2} + 1 & \cdots & 2^{n-1} \end{bmatrix} \quad (4)$$

In the above matrix all the off diagonal elements are same and it can be expressed

$$(2^{n-1} - 2^{n-2} - 1)I_v + 2^{n-2} + 1E_{vv} \quad (5)$$

Hence it proves that the incidence matrix N in (2) gives unequal block sizes, equi-replicated and binary pair wise balanced design with parameters

$$v = n, b = 2^n - n, r = 2^{n-1}, \underline{k} = \left(\underbrace{2, \dots, 2}_{nC_2}; \underbrace{3, \dots, 3}_{nC_3}; \underbrace{n-1, \dots, n-1}_{nC_{n-1}}; n, n \right)' \text{ and } \lambda = 2^{n-2} + 1$$

Numerical Example

Let $n = 4$. So the symmetrical factorial design is 2^4 . The incidence matrix using the above lemma is as follows

$$N = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}_{(4 \times 12)}$$

C-matrix of the above design is obtained as

$$C = \frac{20}{12} \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \\ -1 & -1 & -1 & 3 \end{bmatrix}_{(4 \times 4)} \quad (6)$$

The above matrix can be further simplified as

$$C = \frac{80}{12} \left[I_4 - \frac{1}{4} J_4 J_4^T \right] \quad (7)$$

The matrix in (6) is the C-matrix of the block design with non-zero eigenvalue $\theta = \frac{80}{12}$ with multiplicity 3. The NN^T is

PAIR WISE BALANCED DESIGN

$$NN^T = \begin{bmatrix} 8 & 5 & 5 & 5 \\ 5 & 8 & 5 & 5 \\ 5 & 5 & 8 & 5 \\ 5 & 5 & 5 & 8 \end{bmatrix}_{(4 \times 4)} = 3I_v + 5E_{vv}$$

Obviously this matrix satisfies the conditions of pair wise balanced design. Hence the resulting design is a pair wise balanced design with parameters

$$v = 4, b = 12, r = 8 \text{ and } \underline{k} = \left(\underbrace{2, 2, \dots, 2}_{4 \text{ } C_{2\text{-times}}}; \underbrace{3, 3, \dots, 3}_{4 \text{ } C_{3\text{-times}}}; 4, 4 \right)' \text{ and } \lambda = 5.$$

Pair wise balanced design using 2^n symmetrical factorial design by deleting control and all the main effect treatments

Lemma. Let us consider a 2^n Symmetrical factorial experiment, having 2^n treatment combinations.

Let delete the control treatment as well as all main effects. In result we have $2^n - n - 1$ treatment combinations. Consider this as the blocks of the required design. This we prove with the following example. Let $n = 3$. The $2^3 = 8$ treatment combinations are

$$\begin{array}{ccc} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{array}$$

Delete control treatment. Next delete all treatment combinations whose level of one factor is one while level of other factor is zero. That is, delete all the main effects. Finally, the treatment combinations are

$$\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \end{array}$$

Call this matrix as M which is the combinations of $(2^n - n - 1)$ treatment combinations. Call this as blocks.

Theorem. If there exist a 2^n symmetrical factorial experiment then there always exist un-equal block sizes, equi-replicated, binary pair wise balanced design, by deleting the control treatment and all main effects with the parameters

$$v = n, b = 2^n - n - 1, r = 2^{n-1}, \underline{k} = \left(\underbrace{2, \dots, 2}_{nC_2}, \underbrace{3, \dots, 3}_{nC_3}, \underbrace{n-1, \dots, n-1}_{nC_{n-1}}, n \right)' \text{ and } \lambda = 2^{n-2}$$

Proof: Let us consider a 2^n symmetrical factorial experiment. This has 2^n treatment combinations. Consider n factors as rows and $(2^n - n - 1)$ treatment combinations as blocks. Using the above lemma, we have the following incidence matrix of a design d .

$$N = \begin{bmatrix} 1 & 1 & 0 & 1 & \dots & \dots & 0 & 1 \\ 1 & 0 & 1 & 1 & \dots & \dots & 1 & 1 \\ 0 & 1 & 1 & 1 & \dots & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & \dots & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & \dots & 1 & 1 \\ 0 & 0 & 0 & 0 & \dots & \dots & 1 & 1 \end{bmatrix} \quad (8)$$

For incidence matrix N , among nC_2 columns, element '1' will occur two times and element 0 will occur $(n - 2)$ times. Similarly for nC_3 columns, element '1' will occur three times and element 0 will occur $(n - 3)$ times in each column and so on i.e. for nC_{n-1} columns, element '1' will occur $(n - 1)$ time and element 0 will occur 1 time. More over there will be one more column whose all elements

PAIR WISE BALANCED DESIGN

are unity. Hence numbers of blocks are $b = {}^nC_2 + {}^nC_3 + {}^nC_{n-1} + {}^nC_n$. Since ${}^nC_0 + {}^nC_1 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n$ so $b = (2^n - n - 1)$.

Again each row of N contains $2^{n-1} - 1$ times '1' and n times "zero", so the number of replication is $r = 2^{n-1} - 1$. Since we have considered rows as treatments and hence we have $v = n$ treatments. Similarly columns as blocks and hence $b = 2^n - n - 1$ blocks. Hence

$$\underline{k} = \left(\underbrace{2, \dots, 2}_{{}^nC_2}; \underbrace{3, \dots, 3}_{{}^nC_3}; \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1}}; n \right)'$$

For the design having the incidence matrix N given in equation (8), the NN^T is defined as

$$NN^T = \begin{bmatrix} 2^{n-1}-1 & 2^{n-2} & \dots & 2^{n-2} \\ 2^{n-2} & 2^{n-1}-1 & \dots & 2^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ 2^{n-2} & 2^{n-2} & \dots & 2^{n-1}-1 \end{bmatrix} \quad (9)$$

In the above matrix all the off diagonal elements are same and it can further be expressed as

$$(2^{n-1} - 2^{n-2})I_v + 2^{n-2}E_{vv} \quad (10)$$

Hence it proves that the resulting design is a unequal block sizes, equi-replicated and binary pair wise balanced designs with parameters

$$v = n, b = 2^n - n - 1, r = 2^{n-1} - 1, \underline{k} = \left(\underbrace{2, \dots, 2}_{{}^nC_2}; \underbrace{3, \dots, 3}_{{}^nC_3}; \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1}}; n \right)' \text{ and } \lambda = 2^{n-2}$$

Numerical Example

In a 2^4 factorial design, after deleting the control treatment and all main effects, the incidence matrix is given as

$$N = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Here $v = n = 4$.

The C -matrix is given by

$$C = \begin{bmatrix} 17/4 & -17/12 & -17/12 & -17/12 \\ -17/12 & 17/4 & -17/12 & -17/12 \\ -17/12 & -17/12 & 17/4 & -17/12 \\ -17/12 & -17/12 & -17/12 & 17/4 \end{bmatrix}$$

The non-zero eigenvalues of the above C -matrix is $17/3$ with multiplicity 3.

The NN^T is

$$NN^T = \begin{bmatrix} 7 & 4 & 4 & 4 \\ 4 & 7 & 4 & 4 \\ 4 & 4 & 7 & 4 \\ 4 & 4 & 4 & 7 \end{bmatrix} = 3I_v + 4E_{vv}$$

Hence the resulting design is a pair wise balanced design with parameters

$$v = 4, b = 11, r = 7 \text{ and } \underline{k} = \left(\underbrace{2 \ 2 \ 2 \ 2 \ 2 \ 2}_{4C_2} \underbrace{3 \ 3 \ 3 \ 3}_{4C_3} 4 \right)' \text{ having } \lambda = 4$$

Optimal Pair wise balanced design

A-Optimality: Let d belongs $D(v, b, r, k, \lambda)$ with C_d matrix, where $C_d = R_d - N_d K^{-1} N_d^T$. Design d will be A -Optimal if it maximizes $tr(C_d)$. That is, $tr(C_d) = tr(R_d - N_d K^{-1} N_d^T) = tr(R_d) - tr(N_d K^{-1} N_d^T)$.

For a design d , it can be shown that the sum of the variances of the estimates of all elementary treatment contrast is proportional to the sum of the reciprocals of the non-zero eigenvalues of C .

Let $\theta_1, \theta_2, \theta_3, \dots, \theta_{(v-1)}$ are non-zero eigenvalues. For this design there will be only one non-zero eigenvalues with multiplicities $(v - 1)$ of C_d matrix. That is, $\theta_1 = \theta_2 = \theta_3 = \dots = \theta_{(v-1)} = \theta$ as C -matrix is positive semi-definite. Finally we can say that the design is A -Optimal if

$$\sum_{i=1}^{(v-1)} \frac{1}{\theta_i} \geq \frac{(v-1)^2}{tr(C_d)} \quad (11)$$

D-Optimality: Let $\theta_1, \theta_2, \theta_3, \dots, \theta_{(v-1)}$ are non-zero eigenvalues with multiplicities $(v - 1)$ of C_d matrix of design d . A design is D -Optimal if

$$\prod_{i=1}^{(v-1)} \frac{1}{\theta_i} \leq \prod_{i=1}^{(v-1)} \left(\frac{\sum_{i=1}^{(v-1)} \frac{1}{\theta_i}}{(v-1)} \right) \quad (12)$$

E-Optimality: Let $\theta_1, \theta_2, \theta_3, \dots, \theta_{(v-1)}$ are non-zero eigenvalues with multiplicities $(v - 1)$ of C_d matrix of design d . A design is E -Optimal if

$$Min(\theta_i) \leq \frac{tr(C_d)}{(v-1)} \quad (13)$$

Example: Consider the pair wise balanced design obtained in the numerical example with parameters $v = 4, b = 12, r = 8, k = (2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 4, 4)$. The trace of C -matrix of pair wise balanced design comes out as 20 and non-zero eigenvalue of C -matrix is $\theta = \frac{80}{12}$ with multiplicity 3. Here the inequality (11) holds true which is required condition to be A -Optimal

of pair wise balanced design with equal replication and unequal block sizes and hence pair wise balanced designs is an A -Optimal.

Again the inequalities in (12) and (13) holds true and hence the pair wise balanced design is D -Optimal as well as E -Optimal. Since the constructed pair wise balanced design is A -Optimal, D -Optimal as well as E -Optimal and hence the constructed pair wise balanced design is the universal optimal.

References

Bose, R. C., & Shrikhande, S. S. (1959a). On the falsity of Euler's conjecture about the non-existence of two orthogonal Latin square of order $4t + 2$. *Proceedings of the National Academy of Sciences of the United States of America*, 45(5), 734–737.

Bose, R. C., & Shrikhande, S. S. (1959b). *On the construction of sets of pair wise orthogonal Latin squares and the falsity of conjecture of Euler-II*. Mimeograph Series No. 225, Institute of Statistics, University of North Carolina, Chapel Hill.

Bose, R. C., & Shrikhande, S. S. (1960a). On the composition of balanced incomplete block designs. *Canadian Journal of Mathematics*, 12, 177-188. doi:10.4153/CJM-1960-015-8

Bose, R. C., & Shrikhande, S. S. (1960b). On the construction of sets of mutually orthogonal Latin squares and the falsity of a conjecture of Euler. *Transactions of the American Mathematical Society*, 95(2), 191-209. doi:10.1090/S0002-9947-1960-0111695-3

Effanga, E. O., Ugboh, J. A., Enang, E. I. & Eno, B. E. A. (2009). A tool for constructing pair wise balanced incomplete block design. *Journal of Modern Mathematics and Statistics*, 3(4), 69-72.

Hedayat, A., & Stufken, J. (1989). A relation between Pair wise balanced and variance balanced block designs. *Journal of the American Statistical Association*, 84(407), 753-755. doi:10.1080/01621459.1989.10478830

Some Ridge Regression Estimators and Their Performances

B. M. Golam Kibria

Florida International University
Miami, FL

Shipra Banik

Independent University, Bangladesh
Bashundhara, Dhaka

The estimation of ridge parameter is an important problem in the ridge regression method, which is widely used to solve multicollinearity problem. A comprehensive study on 28 different available estimators and five proposed ridge estimators, KB1, KB2, KB3, KB4, and KB5, is provided. A simulation study was conducted and selected estimators were compared. Some of selected ridge estimators performed well compared to the ordinary least square (OLS) estimator and some existing popular ridge estimators. One of the proposed estimators, KB3, performed the best. Numerical examples were given.

Keywords: Linear regression, mean square error, multicollinearity, ridge regression, simulation study

---

Introduction

Applied researchers are often concerned about models specification under consideration, especially with regards to problems associated with errors. Models specification can be due to omission of one or several relevant variables, inclusion of unnecessary explanatory variables, wrong functional forms, autocorrelation etc. However, for modeling data, there are other problems that also might influence results. This problem occurs in situations when explanatory variables are highly inter-correlated. In practice, there may be strong or near strong linear relationship exist among explanatory variables. Thus, independence assumption of explanatory variables is no longer valid, which causes problem of multicollinearity. In the presence of multicollinearity, the OLS estimator could become unstable due to their large variance, which leads to poor prediction and wrong inference about model parameters. Empirically, problem of multicollinearity can be observed, for example, in cement production, when amount of different compounds in clinkers is

---

Dr. Kibria is a Professor in the Department of Mathematics and Statistics. Email him at: kibriag@fiu.edu. Dr. Banik is an Associate Professor in the Department of Physical Sciences. Email her at: banik@iub.edu.bd.

regressed on the heat evolved of cement (See Muniz and Kibria (2009) for details). Another possible example, when a researcher is interested to predict cholesterol level of patients based on some predictors: age, body weight, blood pressure, food intake and stress causes multicollinearity. In the presence of this noise of the model, regression coefficients may be statistically insignificant or have wrong sign or have large sampling variance that may result in wide confidence interval for individual parameters. With these errors, it is very difficult to make valid statistical inferences and appropriate prediction. Therefore, resolve multicollinearity problem is a serious issue for the linear regression practitioners.

Problem of multicollinearity can be solved by various methods, namely to collect additional data, reselecting variables, principle component regression methods, re-parameterizing the model, ridge regression method, and others. In this paper, we will consider the most widely used ridge regression method. The concept of ridge regression was first proposed by Hoerl and Kennard (1970) to handle multicollinearity problem for engineering data. They found that there is a nonzero value of k (ridge parameter) for which mean squared error (MSE) for the ridge regression estimator is smaller than variance of the ordinary least squares (OLS) estimator. Many authors at different period of times worked in this area of research and developed and proposed different estimators for k . To mention a few, Hoerl and Kennard (1970), Hoerl, Kennard, and Baldwin (1975), McDonald and Galarneau (1975), Lawless and Wang (1976), Dempster, Schatzoff, and Wermuth (1977), Gibbons (1981), Kibria (2003), Khalaf and Shukur (2005), Alkhamisi and Shukur (2008), Muniz and Kibria (2009), Gruber (2010), Muniz, Kibria, Mansson, and Shukur (2012), Mansson, Shukur, and Kibria (2010), and very recently Hefnawy and Farag (2013), Aslam (2014), and Arashi and Valizadeh (2015), among others. Since aforementioned ridge regression estimators are considered by several researchers at different times and under different simulation conditions, they are not comparable as a whole. The objective of this article is to do a comprehensive study on 28 different ridge estimators those are available in literature and compare them based on minimum MSE criterion. Investigation has been carried out using a Monte Carlo simulation. A number of different models have been studied where variance of the random error, correlation among explanatory variables, sample size and unknown coefficient vector were varied. The organization of the paper is as follows. We first review the available methods for estimating k , followed by a Monte Carlo simulation study. Some applications have then been considered and, finally, some concluding remarks are presented.

Statistical Methodology

Ridge Regression Estimators

To describe the ridge regression, consider following multiple linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, $\mathbf{X}$ is an $n \times p$ observed matrix of the regression, and $\mathbf{e}$ is an $n \times 1$ vector of random errors which is distributed as multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}_n$, $\mathbf{I}_n$ being an identity matrix of order n . The OLS estimator of $\boldsymbol{\beta}$ is obtained as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}$$

and covariance matrix of $\hat{\boldsymbol{\beta}}$ is obtained as $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$. It is easy to see that both $\hat{\boldsymbol{\beta}}$ and $\text{Cov}(\hat{\boldsymbol{\beta}})$ are heavily depend on characteristics of the matrix $\mathbf{X}'\mathbf{X}$.

The standard regression model assumes that regressors are nearly independent. However, in many practical situations (e.g. engineering in particular (Hoerl & Kennard, 1970)), often find that regressors are nearly dependent. In that case, the matrix $\mathbf{X}'\mathbf{X}$ becomes ill conditioned (i.e. $\det(\mathbf{X}'\mathbf{X}) \approx 0$). If $\mathbf{X}'\mathbf{X}$ is ill conditioned, then $\hat{\boldsymbol{\beta}}$ is sensitive to a number of errors and therefore meaningful statistical inference becomes very difficult for practitioners. To overcome this problem, Hoerl and Kennard (1970) suggested a small positive number to be added to diagonal elements of the matrix $\mathbf{X}'\mathbf{X}$. Thus resulting estimators are obtained as

$$\begin{aligned} \hat{\boldsymbol{\beta}}(k) &= (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1} \mathbf{X}'\mathbf{y} \\ &= \mathbf{W}\hat{\boldsymbol{\beta}} \end{aligned} \quad (2)$$

where $\mathbf{W} = [\mathbf{I}_p + k\mathbf{C}^{-1}]^{-1}$, $k \geq 0$, $\mathbf{C} = \mathbf{X}'\mathbf{X}$, and $\mathbf{I}_p$ is an identity matrix of order p . This is known as the ridge regression estimator. Since the quantity $[\mathbf{X}'\mathbf{X} + k\mathbf{I}_p]$ in (2) is always invertible, there always exist a unique solution for $\hat{\boldsymbol{\beta}}(k)$. The ridge regression estimator is a biased estimator and, for a positive value of k , this

estimator provides a smaller MSE compared to the OLS estimator. From (2), we observe that as $k \rightarrow 0$, $\hat{\beta}(k) \rightarrow \hat{\beta}$, and as $k \rightarrow \infty$, $\hat{\beta}(k) \rightarrow 0$.

The bias, variance matrix, and MSE expression of $\hat{\beta}(k)$ are respectively given as follows:

$$\begin{aligned} \text{Bias} &= E(\hat{\beta}(k)) - \beta = -kC^{-1}(k)\beta \\ V(\hat{\beta}(k)) &= \sigma^2(WC^{-1}W') \\ \text{MSE}(\hat{\beta}(k)) &= \sigma^2 \text{tr}(WC^{-1}W') + k^2\beta'C^{-2}(k)\beta \end{aligned}$$

where $C(k) = [C + kI_p]$.

The parameter k is known as the “biased” or “ridge” parameter and it must be estimated using real data. Most of recent efforts in the area of multicollinearity and ridge regression estimators have concentrated on estimating the value of k . We will review statistical methodology used to analyze the estimation of k in the next section.

Estimation of Ridge Parameter k

Suppose there exists an orthogonal matrix D such that $D'CD = \Lambda$, where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ contains eigenvalues of the matrix $X'X$. The orthogonal version of (1) is

$$y = X^* \alpha + e \quad (3)$$

where $X^* = XD$ and $\alpha = D'\beta$. Then the generalized ridge regression estimator is given as

$$\hat{\alpha}(k) = (X^{*'}X^* + K)^{-1} X^{*'}y, \quad k \geq 0 \quad (4)$$

where $K = \text{diag}(k_1, k_2, \dots, k_p)$, $k_i > 0$ and $\hat{\alpha}(k) = \Lambda^{-1}X^{*'}y$ is the OLS estimators of α .

It follows from Hoerl and Kennard (1970) that k_i minimizes $\text{MSE}(\hat{\alpha}(k))$, which is defined as

$$MSE(\hat{\alpha}(\hat{k})) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + \hat{k}_i)^2} + \sum_{i=1}^p \frac{\hat{k}_i^2 \hat{\alpha}_i^2}{(\lambda_i + \hat{k}_i)^2} \quad (5)$$

where the λ_i are eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$, $\hat{\alpha}_i$ is the i^{th} element of $\hat{\mathbf{a}}$, and

$$\begin{aligned} \hat{k}_i &= \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n \hat{e}_i^2}{n-p} \\ \hat{e}_i &= y_i - \mathbf{X}_j' \hat{\mathbf{a}}_i \end{aligned}$$

Now we will review available methods in literature to estimate the value of k . Hoerl and Kennard (1970) suggested k to be (denoted here by $\hat{k}_{\text{HK}}$)

$$\hat{k}_{\text{HK}} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\max}^2} \quad (6)$$

where $\hat{\alpha}_{\max}$ is the maximum element of $\hat{\mathbf{a}}$. Hoerl and Kennard claimed that (6) gives smaller MSE than the OLS method.

Hoerl et al. (1975) proposed k to be (denoted here by $\hat{k}_{\text{HKB}}$)

$$\hat{k}_{\text{HKB}} = \frac{p\hat{\sigma}^2}{\hat{\mathbf{a}}'\hat{\mathbf{a}}} \quad (7)$$

Lawless and Wang (1976) suggested k to be (denoted here by $\hat{k}_{\text{LW}}$)

$$\hat{k}_{\text{LW}} = \frac{p\sigma^2}{\hat{\mathbf{a}}'\mathbf{X}'\mathbf{X}\hat{\mathbf{a}}} \quad (8)$$

Hocking, Speed, and Lynn (1979) suggested k to be (denoted here by $\hat{k}_{\text{HSL}}$)

$$\hat{k}_{\text{HSL}} = \hat{\sigma}^2 \frac{\sum_{i=1}^p (\lambda_i \hat{\alpha}_i)^2}{\left(\sum_{i=1}^p \lambda_i \hat{\alpha}_i^2 \right)^2} \quad (9)$$

Kibria (2003) proposed the following estimators for k based on arithmetic mean (AM), geometric mean (GM), and median of $\hat{\sigma}^2 / \hat{\alpha}_i^2$. These are defined as follows:

The estimator based on AM (denoted by $\hat{k}_{\text{AM}}$)

$$\hat{k}_{\text{AM}} = \frac{1}{p} \sum_{i=1}^p \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \quad (10)$$

The estimator based on GM (denoted by $\hat{k}_{\text{GM}}$)

$$\hat{k}_{\text{GM}} = \frac{\hat{\sigma}^2}{\left(\prod_{i=1}^p \hat{\alpha}_i^2 \right)^{\frac{1}{p}}} \quad (11)$$

The estimator based on median (denoted by $\hat{k}_{\text{MED}}$)

$$\text{Median} \left\{ \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \right\}, \quad i = 1, 2, \dots, p \quad (12)$$

Based on modification of $\hat{k}_{\text{HK}}$, Khalaf and Shukur (2005) suggested k to be (denoted by $\hat{k}_{\text{KS}}$)

$$\hat{k}_{\text{KS}} = \frac{\lambda_{\max} \hat{\sigma}^2}{(n-p) \hat{\sigma}^2 + \lambda_{\max} \hat{\alpha}_{\max}^2} \quad (13)$$

where $\lambda_{\max}$ is the maximum eigenvalue of the matrix $\mathbf{X}'\mathbf{X}$.

Following Kibria (2003) and Khalaf and Shukur (2005), Alkhamisi, Khalaf, and Shukur (2006) proposed the following three estimators of k :

$$\hat{k}_{\text{arith}}^{\text{KS}} = \frac{1}{p} \sum_{i=1}^p \frac{\lambda_i \hat{\sigma}_i^2}{(n-p) \hat{\sigma}_i^2 + \lambda_i \hat{\alpha}_i^2} \quad (14)$$

$$\hat{k}_{\text{max}}^{\text{KS}} = \max \left(\frac{\lambda_i \hat{\sigma}_i^2}{(n-p) \hat{\sigma}_i^2 + \lambda_i \hat{\alpha}_i^2} \right) \quad (15)$$

$$\hat{k}_{\text{md}}^{\text{KS}} = \text{median} \left(\frac{\lambda_i \hat{\sigma}_i^2}{(n-p) \hat{\sigma}_i^2 + \lambda_i \hat{\alpha}_i^2} \right) \quad (16)$$

Applying algorithm of GM and square root to Khalaf and Shukur (2005), Kibria (2003), and Alkhamisi et. al (2006), Muniz and Kibria (2009) proposed the following seven estimators of k :

$$\hat{k}_{\text{gm}}^{\text{KS}} = \prod_{i=1}^p \left(\frac{\lambda_i \hat{\sigma}_i^2}{(n-p) \hat{\sigma}_i^2 + \lambda_i \hat{\alpha}_i^2} \right)^{\frac{1}{p}} \quad (17)$$

$$\hat{k}_{\text{KM2}} = \max \left(\frac{1}{\sqrt{\hat{\sigma}^2 / \hat{\alpha}_i^2}} \right) \quad (18)$$

$$\hat{k}_{\text{KM3}} = \max \left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}} \right) \quad (19)$$

$$\hat{k}_{\text{KM4}} = \left(\prod_{i=1}^p \frac{1}{\sqrt{\hat{\sigma}^2 / \hat{\alpha}_i^2}} \right)^{\frac{1}{p}} \quad (20)$$

$$\hat{k}_{\text{KM5}} = \left(\prod_{i=1}^p \frac{\hat{\sigma}^2}{\hat{\alpha}_i^2} \right)^{\frac{1}{p}} \quad (21)$$

$$\hat{k}_{\text{KM6}} = \text{median} \left(\frac{1}{\sqrt{\hat{\sigma}^2 / \hat{\alpha}_i^2}} \right) \quad (22)$$

$$\hat{k}_{\text{KM7}} = \text{median} \left(\sqrt{\frac{\hat{\sigma}^2}{\hat{\alpha}_i^2}} \right) \quad (23)$$

Following Alkhamisi and Shukur (2008) and based square root transformations, Muniz et al. (2012) proposed the following five estimators of k :

$$\hat{k}_{\text{KM8}} = \max \left(\frac{1}{q_i} \right) \quad (24)$$

$$\hat{k}_{\text{KM9}} = \max (q_i) \quad (25)$$

$$\hat{k}_{\text{KM10}} = \left(\prod_{i=1}^p \frac{1}{q_i} \right)^{\frac{1}{p}} \quad (26)$$

$$\hat{k}_{\text{KM11}} = \left(\prod_{i=1}^p q_i \right)^{\frac{1}{p}} \quad (27)$$

$$\hat{k}_{\text{KM12}} = \text{median} \left(\frac{1}{q_i} \right) \quad (28)$$

where $q_i = \frac{\lambda_{\max} \hat{\sigma}^2}{(n-p) \hat{\sigma}^2 + \lambda_{\max} \hat{\alpha}_i^2}$.

Khalaf (2012), based on modification of $\hat{k}_{\text{HK}}$, proposed k to be (denoted by $\hat{k}_{\text{GK}}$)

$$\hat{k}_{\text{GK}} = \hat{k}_{\text{HK}} + \frac{2}{(\lambda_{\max} + \lambda_{\min})}, \quad (29)$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$, respectively.

Nomura (1988) suggested k to be (denoted by $\hat{k}_{\text{HMO}}$)

$$\hat{k}_{\text{HMO}} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^p \left[\hat{\alpha}_i^2 / \left(1 + \left(1 + \lambda_i \left(\frac{\hat{\alpha}_i^2}{\hat{\sigma}^2} \right)^{\frac{1}{2}} \right) \right) \right]} \quad (30)$$

Dorugade and Kashid (2010), based on (7), suggested k to be (denoted by $\hat{k}_{\text{D}}$)

$$\hat{k}_{\text{D}} = \max \left(0, \hat{k}_{\text{HKB}} - \frac{1}{n(\text{VIF}_i)_{\max}} \right) \quad (31)$$

where $\text{VIF}_i = \frac{1}{1-R_i^2}$, $i = 1, 2, \dots, p$ is variance inflation factor of the i^{th} regressor and R_i^2 is the coefficient of determination for the regression of X_i on other covariates, $X_1, X_2, \dots, X_i, X_{i+1}, \dots, X_p$ (a regression equation without response variable).

Crouse, Jin, and Hanumara (1995), for $k > 0$ and using unbiased ridge regression (URR) estimator $(k, \mathbf{J}) = (\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}(\mathbf{X}'\mathbf{y} + \mathbf{J}k)$, $k \geq 0$, where $\mathbf{J} \sim \left(\boldsymbol{\beta}, \frac{\sigma^2}{k} \right)$, proposed k to be (denoted by $\hat{k}_{\text{CJH}}$)

$$\hat{k}_{\text{CJH}} = \begin{cases} \frac{p\hat{\sigma}^2}{(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J})'(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J})} & \text{if } (\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J})'(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J}) > \hat{\sigma}^2 \text{tr}(\mathbf{X}'\mathbf{X})^{-1} \\ \frac{p\hat{\sigma}^2}{(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J})'(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \mathbf{J}) - \hat{\sigma}^2 \text{tr}(\mathbf{X}'\mathbf{X})^{-1}} & \text{otherwise} \end{cases} \quad (32)$$

Batah and Gore (2009), using modified URR (known as MUR) estimator for $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta}_{\text{J}}(k) = \left[\mathbf{I} - k(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1} \right] \left[(\mathbf{X}'\mathbf{X} + k\mathbf{I}_p)^{-1}(\mathbf{X}'\mathbf{y} + k\mathbf{J}) \right],$$

suggested k to be (denoted by $\hat{k}_{\text{FG}}$)

$$\hat{k}_{FG} = \frac{p\hat{\sigma}^2}{\sum_{i=1}^p \left[\hat{\alpha}_i^2 / \left(\frac{\hat{\alpha}_i^4 \lambda_i^2}{4\hat{\sigma}^4} + \frac{6\hat{\alpha}_i^2 \lambda_i}{\hat{\sigma}^2} \right)^{\frac{1}{2}} - \frac{6\hat{\alpha}_i^2 \lambda_i}{\hat{\sigma}^2} \right]} \quad (33)$$

In the next section, we evaluated 28 different ridge estimators that are defined in equations (6) to (33) to know which estimators show better performances under our simulation study flowchart.

The Monte Carlo Simulation

The aim of this study is to compare the performance of different ridge estimators and find some good estimators for practitioners. Because a theoretical comparison is not possible, a simulation study has been conducted using MATLAB 8.0. The design of this simulation study depends on what factors are expected to affect properties of estimators under investigation and what criteria are being used to judge results. Because the degree of collinearity among explanatory variables (X s) is of central importance, we followed Kibria (2003) in generating X s using the following equation:

$$X_{ij} = (1 - \gamma^2)^{\frac{1}{2}} z_{ij} + \gamma z_{ip}, i = 1, 2, \dots, n, j = 1, 2, \dots, p \quad (34)$$

where z_{ij} are independent standard normal pseudo-random numbers and γ represents correlation between any two X s. These variables are standardized so that $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$ are in correlation forms. The n observations for $\mathbf{y}$ are determined by the following equation:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + e_i, i = 1, 2, \dots, n \quad (35)$$

where the e_i are i.i.d. $N(0, \sigma^2)$ and, without loss of any generality, we will assume zero intercept for (35).

Correlation Coefficient, Sample Size, and Replications

A number of factors such as γ , n , σ , and number of replications can affect properties of the estimators. Since our objective is to compare performance of estimators according to the strength of multicollinearity, we used different degrees of

correlation between variables and let $\gamma=0.70, 0.80$, and 0.90 . Eigenvalues and eigenvectors of the correlation matrix indicate the degree of multicollinearity. One of the possible widely used estimators to measure the strength of multicollinearity called condition number (Vinod & Uallh, 1981) is defined as follows

$$\kappa = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \quad (36)$$

where $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and the smallest eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$, respectively. If $\lambda_{\min} = 0$, then κ is infinite, which means perfect multicollinearity among X s. If $\lambda_{\max} = \lambda_{\min}$, then $\kappa = 1$ and the X s are said to be orthogonal. Large values of κ indicate serious multicollinearity. Usually, a κ between 30 and 100 indicates a moderate to strong correlation, and a κ greater than 100 suggests severe multicollinearity. An eigenvalue that approaches 0 indicates a very strong linear dependency between X s.

Because a purpose of the study is to see the effect of n on the performance of the estimators, $n = 20$ and $n = 50$ were considered. The number of X s is also of great importance since the bad impact of the collinearity on MSE might be stronger when there are more X s in the model. Also, $p = 5$ is used in our study. To see whether the magnitude of σ has a significant effect on the performance of the proposed estimators, we used $\sigma = 0.01, 0.5, 1.0$, and 5.0 . For each set of X s, we selected coefficients $\beta_1, \beta_2, \dots, \beta_p$ as normalized eigenvectors corresponding to the largest eigenvalue of the matrix $\mathbf{X}'\mathbf{X}$ subject to constraint $\boldsymbol{\beta}'\boldsymbol{\beta} = 1$. Thus, for $n, p, \boldsymbol{\beta}, \lambda, \gamma$, and σ , sets of X s are generated. Then the experiment was repeated 5000 times by generating new error terms. Values of k of different selected estimators and average MSEs are estimated and presented them in Tables 5 to 10. In these tables, average k was calculated for ridge estimators and the proportion of replications for which OLS estimators produce a smaller MSE than selected ridge regression estimators and are presented in parenthesis.

Results

Performance as a Function of σ

In Tables 5 to 10, the MSEs of selected estimators are provided as a function of σ . To understand very clearly for $\gamma = 0.70$ and $n = 20$, performance of estimators as a function of σ is provided in Figure 1. From results, we observed as σ increases,

MSEs also increases. Also for smaller σ (e.g. $\sigma = 0.1$), performances of selected estimators do not differ greatly. It is noticeable that all ridge estimators have smaller MSE than the OLS estimator except $\sigma = 0.1$. The performance of the GM, KM2, KM3, KM4, KM5, KM6, KM7, KM8, KM9, KM10, KM11, HMO, and FG estimators are better compared to the rest of estimators.

Figure 1. Performance of estimators as a function of σ

However, when σ is large (e.g. $\sigma = 5.0$), the GM, MED, KM3, HMO, CJH, and FG estimators outperform all other estimators in the sense of smaller MSE (see Figure 1). A significant increase in MSEs were observed when a shifting from $\sigma = 1.0$ to $\sigma = 5.0$.

Performance as a Function of γ

MSEs of selected estimators were also analyzed as a function of γ for selected values of n , p , and σ . These results are available on request from the authors. For a clear understanding, for $(\sigma = 1, n = 20)$ and $(\sigma = 5, n = 50)$, performances of estimators are provided in Figure 2 and Figure 3, respectively. It is clear that, as γ increases, the MSEs also increase (see Figures 2 and 3). When γ increases (see Figure 3), higher correlation between X s resulted in an increase of MSEs of ridge estimators. In general, HSL, GM, MED, KS_Max, KM2, KM3, KM5, KM8, KM9, HMO, and FG performed better than other estimators.

Performance as a Function of n

MSEs of selected estimators were evaluated as a function of n , for which tabulated results are available from the authors on request. For given $\gamma = 0.8$, $p = 5$, performances of estimators as a function of n for $\sigma = 1$ and $\sigma = 5$ are provided in Figure 4 and Figure 5, respectively. We observed that, as n increases, MSEs decrease and the performance of estimators do not vary significantly. An important change has been observed in MSEs when σ shifts from 1 to 5. We observed that, in general, when n increases, MSEs decrease, which is true for large values of γ and σ . Performance of estimators does not vary greatly for small values of σ and γ .

Figure 2. Performance of estimators as a function of γ for $\sigma = 1$ and $n = 20$

Figure 3. Performance of estimators as a function of γ for $\sigma = 5$ and $n = 50$

Figure 4. Performance of estimators as a function of n for $\gamma = 0.8$ and $\sigma = 1.0$

Figure 5. Performance of estimators as a function of n for $\gamma = 0.8$ and $\sigma = 5.0$

Some Proposed Ridge Estimators

Based on the above, the following five new estimators of k are proposed:

1. KB1 = Arithmetic mean of (GM, MED, KM3, HMO, CJH, FG)
2. KB2 = Median(GM, MED, KM3, HMO, CJH, FG)
3. KB3 = Max(GM, MED, KM3, HMO, CJH, FG)
4. KB4 = Geometric mean of (GM, MED, KM3, HMO, CJH, FG)
5. KB5 = Harmonic mean of (GM, MED, KM3, HMO, CJH, FG)

MSEs values for $n = 10, 20$, and 30 , $\gamma = 0.9$, and $p = 5$ are reported for $\sigma = 3$ and $\sigma = 10$ in [Table A7](#) and [Table A8](#), respectively, for 28 selected existing estimators and our proposed 5 ridge estimators. For better understanding, MSEs are plotted in [Figures 6](#) and [7](#). It appears from these results that all proposed estimators are performing well under some conditions. However, proposed KB3 performed the best followed by KB1 (See [Figures 6](#) and [7](#)).

Figure 6. Performance of estimators as a function of n for $\gamma = 0.9$ and $\sigma = 3.0$

Figure 7. Performance of estimators as a function of n for $\gamma = 0.9$ and $\sigma = 10.0$

Application

Example 1

Consider an example which has been taken from Pasha and Shah (2004) to compare the performances of the selected estimators. The following regression model is considered:

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + e_i, i = 1, 2, \dots, n \quad (37)$$

where y_i = number of persons employed (million), X_{i1} = land cultivated (million hectares), X_{i2} = inflation rate (%), X_{i3} = number of establishments, X_{i4} = population (million), X_{i5} = literacy rate (%), and $n=28$. For details about the data set, see Pasha and Shah (2004).

Table 1. Correclations among exclamatory variables

	X_{i1}	X_{i2}	X_{i3}	X_{i4}	X_{i5}	y_i
X_{i1}	1.0000	0.6573	0.9427	0.9761	0.9564	0.9731
X_{i2}	0.6573	1.0000	0.6232	0.7062	0.6905	0.6926
X_{i3}	0.9427	0.6232	1.0000	0.9633	0.8672	0.9437
X_{i4}	0.9761	0.7062	0.9633	1.0000	0.9506	0.9930
X_{i5}	0.9564	0.6905	0.8672	0.9506	1.0000	0.9572
y_i	0.9731	0.6926	0.9437	0.9930	0.9572	1.0000

Figure 8. MSE of selected ridge estimators

The correlation matrix of X s in (37) is presented in Table 1. It is observed that the X s are highly correlated. Moreover, $\kappa = 38115.32$, which implies the existence of multicollinearity in the data set. So it is adequate to compare proposed ridge estimators with the real data set. Estimated MSEs along with ridge regression coefficients are presented in Table 2 and, for a better presentation, MSEs are plotted in Figure 8.

The MSE of estimators is estimated by

$$\text{MSE}(\hat{\beta}) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + \hat{k})^2} + \sum_{i=1}^p \frac{\hat{k}_i \beta_i^2}{(\lambda_i + \hat{k}_i)^2} \quad (38)$$

where $\hat{k}$ is one of $\hat{k}_{\text{HK}}, \hat{k}_{\text{HKB}}, \dots, \hat{k}_{\text{KB5}}$, and other terms are explained in (5). It is evident from Table 2 and Figure 8 that all ridge estimators perform better than the OLS estimator. However, HKB, AM, KM4, KM6, KM10, KM12, KD, and our five proposed estimators are performing better as compared to other ridge estimators.

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table 2. MSE and estimated ridge regression coefficients of the estimators

Estimators	MSE	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
OLS	1.6578	-1.2600	0.3123	-0.0623	-0.2276	0.0068
HK	1.3726	-0.0969	0.2812	-0.0604	-0.2272	0.0069
HKB	1.0182	-0.2889	0.3029	-0.0618	-0.2275	0.0069
LW	1.1841	-1.2598	0.3123	-0.0623	-0.2276	0.0069
HSL	1.3180	-0.0001	0.0013	-0.0009	-0.0547	0.0068
AM	1.0182	-0.2889	0.3029	-0.0618	-0.2275	0.0068
GM	1.5638	-0.0181	0.1913	-0.0526	-0.2256	0.0067
MED	1.4714	-0.0534	0.2584	-0.0588	-0.2269	0.0066
KS	1.3726	-0.0969	0.2812	-0.0604	-0.2272	0.0063
KS_AM	1.5918	-0.0003	0.0066	-0.0043	-0.1394	0.0064
KS_MAX	1.4179	-0.0001	0.0010	-0.0009	-0.0547	0.0068
KS_MED	1.4164	-0.0772	0.2737	-0.0599	-0.2271	0.0066
KS_GM	1.4258	-0.0730	0.2716	-0.0597	-0.2271	0.0068
KM2	1.1505	-0.2088	0.2984	-0.0615	-0.2274	0.0060
KM3	1.6268	-0.0049	0.0924	-0.0368	-0.2204	0.0068
KM4	0.7061	-0.8279	0.3108	-0.0622	-0.2275	0.0059
KM5	1.3945	-0.0870	0.2777	-0.0602	-0.2272	0.0068
KM6	0.6951	-0.6600	0.3097	-0.0622	-0.2275	0.0066
KM7	1.2737	-0.1441	0.2915	-0.0611	-0.2273	0.0063
KM8	1.3244	-0.1194	0.2870	-0.0608	-0.2273	0.0068
KM9	1.4179	-0.0001	0.0013	-0.0009	-0.0547	0.0067
KM10	0.9905	-1.1420	0.3120	-0.0623	-0.2275	0.0066
KM11	1.5638	-0.0181	0.1913	-0.0526	-0.2256	0.0069
KM12	0.7842	-0.9608	0.3114	-0.0623	-0.2275	0.0068
GK	1.3726	-0.0969	0.2812	-0.0604	-0.2272	0.0068
HMO	1.2170	-0.1729	0.2952	-0.0613	-0.2274	0.0070
KD	1.0182	-0.2889	0.3029	-0.0618	-0.2275	0.0062
CJH	1.4740	-0.0523	0.2574	-0.0587	-0.2269	0.0064
FG	1.2795	-0.1412	0.2910	-0.0610	-0.2273	0.0069
KB1	0.9902	-0.0193	0.1959	-0.0531	-0.2257	0.0065
KB2	0.8727	-0.0528	0.2579	-0.0587	-0.2269	0.0068
KB3	0.9296	-0.0049	0.0926	-0.0368	-0.2204	0.0066
KB4	0.9045	-0.0438	0.2486	-0.0580	-0.2268	0.0062
KB5	0.9165	-0.0771	0.2736	-0.0599	-0.2271	0.0068

Example 2

Consider the data set on total national research and development expenditures as a percent of gross national product originally due to Gruber (1998) and later by Akdeniz and Erol (2003), among others. The regression model is defined as

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + e_i, i = 1, 2, \dots, n \quad (39)$$

where y = percent spent by United States, X_1 = percent spent by France, X_2 = percent spent by West Germany, X_3 = percent spent by Japan, and X_4 = percent spent by the Soviet Union. The correlation matrix of X s in (39) is tabulated in Table 3. We found that the X s are highly correlated. Moreover, $\kappa = 93.6823$ implies the existence of multicollinearity in the data set so it is reasonable to evaluate proposed ridge estimators with the real data set. Estimated MSEs along with regression coefficients are tabulated in Table 4 and, for a better presentation, MSEs are presented in Figure 9. It is evident from Table 4 and Figure 9 that all ridge estimators outperformed the OLS estimator. However, all ridge estimators except KM2, KM3, KM4, KM5, KM6, KM7, KM8, KM10, and KM12 have smaller MSE than the OLS estimator.

Table 3. Correlations among the variables.

	X_{i1}	X_{i2}	X_{i3}	X_{i4}	y_i
X_{i1}	1.0000	0.8877	0.9248	0.3090	0.9776
X_{i2}	0.8877	1.0000	0.9621	0.1573	0.9080
X_{i3}	0.9248	0.9621	1.0000	0.3276	0.9565
X_{i4}	0.3090	0.1573	0.3276	1.0000	0.3482
y_i	0.9776	0.9080	0.9565	0.3482	1.0000

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table 4. MSEs and the estimated ridge regression coefficients of the estimators

Estimators	MSE	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	1.2595	-0.1623	0.4616	0.1733	0.4462
HK	0.0628	-0.1343	0.3975	0.1716	0.4462
HKB	0.0588	-0.1177	0.3569	0.1702	0.4462
LW	0.0804	-0.1618	0.4605	0.1733	0.4462
HSL	0.0622	-0.1327	0.3937	0.1715	0.4462
AM	0.0588	-0.1177	0.3569	0.1702	0.4462
GM	0.0602	-0.1026	0.3181	0.1686	0.4462
MED	0.0588	-0.1170	0.3534	0.1702	0.4462
KS	0.0628	-0.1344	0.3975	0.1716	0.4462
KS_AM	0.0588	-0.1176	0.3566	0.1702	0.4462
KS_MAX	0.0724	-0.0798	0.2557	0.1652	0.4462
KS_MED	0.0638	-0.1367	0.4031	0.1718	0.4462
KS_GM	0.0613	-0.1302	0.3875	0.1713	0.4462
KM2	0.2646	-0.0005	0.0017	0.0101	0.4295
KM3	0.1765	-0.0203	0.0720	0.1302	0.4459
KM4	0.2613	-0.0008	0.0029	0.0162	0.4361
KM5	0.1463	-0.0315	0.1104	0.1451	0.4460
KM6	0.2621	-0.0007	0.0026	0.0146	0.4349
KM7	0.1523	-0.0293	0.1022	0.1426	0.4460
KM8	0.2887	-0.0006	0.0001	0.0009	0.3057
KM9	0.0931	-0.0605	0.2005	0.1605	0.4462
KM10	0.2721	-0.0001	0.0004	0.0025	0.3833
KM11	0.0602	-0.1026	0.3181	0.1686	0.4462
KM12	0.2764	-0.0001	0.0003	0.0017	0.3578
GK	0.0588	-0.1160	0.3525	0.1701	0.4462
HMO	0.0726	-0.0794	0.2552	0.1651	0.4462
KD	0.0643	-0.1377	0.4056	0.1718	0.4462
CJH	0.0713	-0.0810	0.2597	0.1654	0.4462
FG	0.0896	-0.0632	0.2084	0.1613	0.4462
KB1	0.0705	-0.0554	0.1851	0.1588	0.4461
KB2	0.0819	-0.0802	0.2574	0.1653	0.4462
KB3	0.0765	-0.0203	0.0720	0.1302	0.4459
KB4	0.0843	-0.0747	0.2420	0.1642	0.4460
KB5	0.0867	-0.0876	0.2780	0.1666	0.4462

Figure 9. MSE of selected ridge estimators.

Conclusions

Based on our simulation results, the following conclusions can be drawn: As σ increases, MSE have a negative effect, meaning that MSE increases. As γ increases, MSE also increases. When n increases, MSE decreases even when γ and σ are large. In all situations, all ridge estimators have smaller MSE than the OLS estimator. When $\sigma = 5.0$, GM, KM3, MED, KMO, CJH, and FG outperformed all other estimators in the sense of producing smaller MSE. Two real life examples have been studied. Based on the results of simulations and numerical examples, estimators HSL, AM, GM, MED, KS_MAX, KM2, KM3, KM5, KM8, KM9, KMO, CJH, FG, and proposed KB1, KB2, KB3, KB4, and KB5 performed better than the rest in the sense of small MSE and may be recommended to practitioners.

Acknowledgements

This paper is dedicated to all who sacrificed themselves during the liberation war that started on March 26, 1971 and ended on December 16, 1971 to bring out the freedom of our beautiful Bangladesh.

References

- Akdeniz, F. & Erol, H. (2003). Mean squared error matrix comparisons of some biased estimators in linear regression. *Communications in Statistics – Theory and Methods*, 32(12), 2389-2413. doi: [10.1081/STA-120025385](https://doi.org/10.1081/STA-120025385)
- Alkhamisi, M., Khalaf, G., & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics – Theory and Methods*, 35(11), 2005-2020. doi: [10.1080/03610920600762905](https://doi.org/10.1080/03610920600762905)
- Alkhamisi, M. & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics – Theory and Methods*, 37(4), 544-564. doi: [10.1080/03610920701469152](https://doi.org/10.1080/03610920701469152)
- Arashi, M. & Valizadeh, T. (2015). Performance of Kibria's methods in partial linear ridge regression model. *Statistical Papers*, 56(1), 231-246. doi: [10.1007/s00362-014-0578-6](https://doi.org/10.1007/s00362-014-0578-6)
- Aslam, M. (2014). Performance of Kibria's method for the heteroscedastic ridge regression model: Some Monte Carlo evidence. *Communications in Statistics – Simulation and Computation*. 43(4), 673-686. doi: [10.1080/03610918.2012.712185](https://doi.org/10.1080/03610918.2012.712185)
- Batah, F. S. M., & Gore, S. D. (2009). Ridge regression estimator: Combining unbiased and ordinary ridge regression methods of estimation. *Surveys in Mathematics and its Applications*, 4, 99-109.
- Crouse, R., Jin, C., & Hanumara, R. (1995). Unbiased ridge estimation with prior informatics and ridge trace. *Communications in Statistics – Theory and Materials*, 24(9), 2341-2354. doi: [10.1080/03610929508831620](https://doi.org/10.1080/03610929508831620)
- Dempster, A. P., Schatzoff, M., & Wermuth, N. (1977). A simulation study of alternatives to ordinary least squares. *Journal of the American Statistical Association*, 72(357), 77-91. doi: [10.1080/01621459.1977.10479910](https://doi.org/10.1080/01621459.1977.10479910)
- Dorugade, A. V. & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences*, 4(9), 447-456.
- Gibbons, D. G. (1981). A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373), 131-139. doi: [10.1080/01621459.1981.10477619](https://doi.org/10.1080/01621459.1981.10477619)
- Gruber, M. H. J. (1998). *Improving efficiency by shrinkage the James-Stein and ridge regression estimators*. New York, NY: Marcel Dekker.
- Gruber, M. H. J. (2010). *Regression estimators* (2nd Ed.). Baltimore, MD: Johns Hopkins University Press.

- Hefnawy, E. A. & Farag A. (2013). A combined nonlinear programming model and Kibria method for choosing ridge parameter regression. *Communications in Statistics – Simulation and Computation*, 43(6), . doi: 10.1080/03610918.2012.735317
- Hocking, R. R., Speed, F. M., & Lynn, M. J. (1976). A class of biased estimators in linear regression. *Technometrics*, 18(4), 55-67. doi: 10.1080/00401706.1976.10489474
- Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for non-orthogonal problems. *Technometrics*, 12(1), 55-67. doi: 10.1080/00401706.1970.10488634
- Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics*, 4(2), 105-123. doi: 10.1080/03610927508827232
- Khalaf, G. (2012). A proposed ridge parameter to improve the least squares estimator. *Journal of Modern Applied Statistical Methods*, 11(2), 443-449.
- Khalaf, G. & Shukur, G. (2005). Choosing ridge parameters for regression problems. *Communications in Statistics – Theory and Methods*, 34(5), 1177-1182. doi: 10.1081/STA-200056836
- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics – Simulation and Computation*, 32(2), 419-435. doi: 10.1081/SAC-120017499
- Lawless, J. F. & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics – Theory and Methods*, 5(4), 307-323. doi: 10.1080/03610927608827353
- Mansson, K., Shukur, G. & Kibria, B. M. G. (2010). On some ridge regression estimators: A Monte Carlo simulation study under different error variances. *Journal of Statistics*, 17(1), 1-22.
- McDonald, G. C. & Galarneau, D. I. (1975). A Monte Carlo evaluation of ridge-type estimators. *Journal of the American Statistical Association*, 70(350), 407-416. doi: 10.1080/01621459.1975.10479882
- Muniz, G. & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparison. *Communications in Statistics – Simulation and Computation*, 38(3), 621-630. doi: 10.1080/03610910802592838
- Muniz, G., Kibria, B. M. G., Mansson, K., & Shukur, G. (2012). On developing ridge regression parameters: A graphical investigation. *Statistics and Operations Research Transactions*, 36(2), 115-138.

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Nomura, M. (1988). On the almost unbiased ridge regression estimation. *Communication in Statistics – Simulation and Computation*, 17(3), 729-743. doi: [10.1080/03610918808812690](https://doi.org/10.1080/03610918808812690)

Pasha, G. R. & Shah, M. A. A. (2004). Application of ridge regression to multicollinear data. *Journal of Research (Science)*, 15(1), 97-106.

Vinod, H. D. & Ullah, A. (1981). *Recent advances in regression models*. New York, NY: Marcel Dekker.

Appendix

Table A1. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 20$, $p = 5$, and $\gamma = 0.7$. Condition number $\kappa = 26.53$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.8001	1.0294	1.8532	27.9287
HK	0.8017 (0.0275, 96.3)	0.9957 (0.5307, 37.6)	1.4446 (1.5525, 18.80)	11.3570 (82.1698, 0.08)
HKB	0.8031 (0.0504, 96.3)	0.9720 (1.1334, 39.6)	1.2765 (3.4790, 20.60)	8.7257 (9.7970, 0.08)
LW	0.8023 (0.0025, 96.3)	1.0218 (0.0620, 35.3)	1.7174 (0.2358, 16.60)	17.8015 (1.4258, 0.16)
HSL	0.8018 (0.0282, 96.3)	0.9759 (0.7261, 39.5)	1.2452 (2.9035, 21.60)	7.7684 (12.0529, 0.08)
AM	0.8031 (0.0504, 96.3)	0.9720 (1.1334, 39.6)	1.2765 (3.4790, 20.60)	8.7257 (9.7970, 0.08)
GM	0.8036 (0.0605, 96.3)	0.9488 (2.3970, 41.4)	1.1223 (12.4590, 20.60)	3.2072 (46.3951, 0.08)
MED	0.8034 (0.0582, 96.3)	0.9600 (1.7633, 40.5)	1.1681 (10.4330, 22.30)	3.8485 (43.8037, 0.08)
KS	0.8017 (0.0272, 96.3)	0.9985 (0.4358, 37.3)	1.5160 (0.9018, 21.80)	15.1509 (1.5806, 0.20)
KS_AM	0.8035 (0.0588, 96.3)	0.9904 (0.4064, 37.8)	1.5265 (0.6275, 18.40)	16.2535 (0.9005, 0.20)
KS_MAX	0.8076 (0.1326, 96.3)	0.9708 (0.7990, 39.8)	1.3412 (1.4414, 17.70)	8.9894 (2.6993, 0.20)
KS_MED	0.8026 (0.0435, 96.3)	0.9922 (0.3735, 37.6)	1.5860 (0.4658, 19.80)	19.9875 (0.4988, 0.16)
KS_GM	0.8030 (0.0490, 96.3)	0.9975 (0.3156, 37.4)	1.6066 (0.4259, 17.40)	20.005 (0.5052, 0.16)
KM2	0.9044 (6.3585, 96.3)	0.9481 (1.5384, 43.7)	1.3699 (1.0418, 17.20)	14.8189 (0.8429, 0.12)
KM3	0.8200 (0.3931, 92.6)	0.9308 (9.4316, 45.0)	1.0795 (217.0360, 18.50)	3.0143 (76.7550, 0.08)
KM4	0.8922 (4.3038, 95.9)	0.9680 (0.7983, 40.6)	1.5727 (0.4285, 22.40)	22.8690 (0.2390, 0.12)
KM5	0.8131 (0.2414, 92.9)	0.9550 (1.4246, 41.2)	1.2430 (2.9300, 17.60)	7.2046 (5.5596, 0.16)
KM6	0.8934 (4.4212, 96.1)	0.9629 (0.9310, 41.3)	1.5454 (0.5069, 20.70)	22.5850 (0.2727, 0.12)
KM7	0.8128 (0.2366, 92.8)	0.9608 (0.2220, 40.8)	1.2757 (2.4513, 17.48)	7.6965 (5.0834, 0.16)
KM8	0.9338 (42.3390, 96.1)	0.9416 (12.9320, 46.8)	1.2885 (1.6238, 20.60)	12.5045 (1.2550, 0.12)
KM9	0.8083 (0.1521, 92.2)	0.9422 (2.1465, 43.6)	1.2042 (2.6788, 18.90)	8.5190 (2.9051, 0.20)
KM10	0.9277 (19.6900, 96.3)	0.9556 (1.2124, 42.5)	1.4762 (0.6735, 21.00)	19.0340 (0.5044, 0.12)
KM11	0.8035 (0.0587, 92.5)	0.9704 (0.9319, 39.5)	1.3472 (1.6037, 17.78)	11.6359 (2.0797, 0.20)
KM12	0.9287 (20.7930, 96.3)	0.9554 (1.3005, 42.5)	1.5019 (0.6313, 19.60)	20.5370 (0.4299, 0.16)
GK	0.8042 (0.0705, 96.3)	0.9923 (0.5737, 37.7)	1.4313 (1.5955, 17.60)	11.1205 (82.2120, 0.08)
HMO	0.8456 (1.2774, 95.2)	0.9255 (8.9203, 43.7)	0.9871 (18.1095, 22.80)	2.7934 (28.6862, 0.20)
KD	0.7967 (0.0077, 45.2)	1.0030 (1.0936, 38.8)	1.2757 (3.4679, 20.72)	8.0586 (9.8342, 0.16)
CJH	0.8338 (0.7991, 96.4)	0.9738 (24.1710, 45.8)	1.0909 (23.3341, 23.80)	4.7328 (28.8315, 0.20)
FG	0.8116 (0.2832, 96.5)	0.9483 (4.2218, 42.5)	1.0504 (9.0792, 22.64)	3.7383 (15.1262, 0.16)

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table A2. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 20$, $p = 5$, and $\gamma = 0.8$. Condition number $\kappa = 44.43$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.8456	1.2497	2.3874	39.8284
HK	0.8489 (0.0293, 97.8)	1.1623 (0.5070, 34.4)	1.6979 (1.4373, 14.6)	15.8668 (539.9400, 0.08)
HKB	0.8513 (0.0500, 97.8)	1.1153 (1.0862, 36.6)	1.4686 (3.0389, 15.9)	11.6959 (7.2387, 0.12)
LW	0.8459 (0.0026, 97.8)	1.2279 (0.0647, 32.2)	2.0825 (0.2464, 13.2)	21.2712 (1.4494, 0.08)
HSL	0.8490 (0.0299, 97.7)	1.1160 (0.7767, 37.0)	1.3292 (3.1169, 17.0)	9.1223 (11.3958, 0.16)
AM	0.8513 (0.0500, 97.8)	1.1153 (1.0862, 36.6)	1.4686 (3.0389, 15.9)	11.6959 (7.2387, 0.12)
GM	0.8520 (0.0571, 97.7)	1.0750 (2.5046, 38.2)	1.2290 (9.8068, 17.0)	3.7616 (36.1258, 0.16)
MED	0.8518 (0.0547, 97.8)	1.0927 (1.8202, 37.2)	1.2786 (7.0498, 16.6)	5.0127 (33.3226, 0.16)
KS	0.8489 (0.0290, 97.8)	1.1694 (0.4212, 34.2)	1.7861 (0.8497, 14.4)	19.5872 (1.5105, 0.08)
KS_AM	0.8512 (0.0494, 97.8)	1.1639 (0.3346, 34.6)	1.7905 (0.5518, 13.8)	19.4642 (0.8356, 0.08)
KS_MAX	0.8565 (0.0987, 97.6)	1.1211 (0.6671, 36.9)	1.4515 (1.5492, 15.9)	8.7602 (2.9129, 0.08)
KS_MED	0.8500 (0.0380, 97.8)	1.1720 (0.2890, 34.1)	1.9516 (0.3353, 13.6)	27.6417 (0.3535, 0.04)
KS_GM	0.8506 (0.0437, 97.8)	1.1816 (0.2461, 33.9)	1.9671 (0.3243, 13.5)	27.1016 (0.3818, 0.04)
KM2	0.9848 (6.1657, 93.6)	1.0671 (1.6257, 40.6)	1.4754 (1.1924, 15.0)	15.5751 (0.9925, 0.04)
KM3	0.8786 (0.3496, 96.9)	1.0371 (10.4610, 40.2)	1.1570 (29.6414, 17.7)	3.0807 (114.9800, 0.12)
KM4	0.9756 (4.4381, 93.8)	1.1038 (0.8180, 37.4)	1.8091 (0.4621, 13.8)	28.7349 (0.2695, 0.04)
KM5	0.8697 (0.2345, 97.2)	1.0834 (1.4215, 38.2)	1.3671 (2.6928, 16.6)	8.0125 (4.9376, 0.08)
KM6	0.9766 (4.5711, 93.7)	1.0946 (0.9472, 38.2)	1.7758 (0.5336, 14.2)	27.9144 (0.3165, 0.04)
KM7	0.8693 (0.2289, 97.2)	1.0927 (1.2106, 37.6)	1.3998 (2.2861, 16.1)	8.9963 (4.4526, 0.08)
KM8	0.9855 (39.8268, 89.9)	1.0499 (3.2641, 43.2)	1.3653 (2.0516, 15.7)	12.5204 (1.6050, 0.04)
KM9	0.8582 (0.1235, 97.2)	1.0554 (2.2557, 39.6)	1.2844 (2.8385, 16.6)	8.3068 (3.1120, 0.08)
KM10	0.9948 (20.8949, 92.1)	1.0804 (1.2369, 39.4)	1.6549 (0.6999, 14.3)	22.3960 (0.5245, 0.04)
KM11	0.8519 (0.0558, 97.7)	1.1111 (0.9332, 36.9)	1.4979 (1.5784, 15.6)	13.0713 (2.0425, 0.08)
KM12	0.9946 (22.2324, 92.0)	1.0793 (1.3071, 39.7)	1.7157 (0.6401, 14.3)	25.1129 (0.4413, 0.04)
GK	0.8534 (0.0697, 97.6)	1.1548 (0.5474, 34.8)	1.6721 (1.4778, 14.6)	15.3714 (539.9800, 0.08)
HMO	0.9096 (0.8998, 95.6)	1.0082 (6.6626, 39.4)	1.1057 (12.5907, 17.1)	4.4618 (18.1221, 0.16)
KD	0.8465 (0.0077, 45.3)	1.1209 (1.0368, 36.6)	1.4844 (2.9895, 15.9)	12.0123 (7.1893, 0.12)
CJH	0.9308 (2.3565, 96.6)	1.0571 (17.7803, 41.5)	1.2212 (15.4247, 17.4)	6.8747 (19.0881, 0.16)
FG	0.8732 (0.2763, 97.2)	1.0367 (3.6773, 39.4)	1.1695 (7.1088, 17.1)	5.3217 (10.4628, 0.12)

Table A3. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 20$, $p = 5$, and $\gamma = 0.9$. Condition number $\kappa = 99.57$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.9252	1.7372	4.072	78.5675
HK	0.9337 (0.0312, 93.84)	1.4437 (0.4401, 25.60)	2.3761 (0.4630, 8.16)	29.8253 (1.0378, 0.08)
HKB	0.9383 (0.0498, 93.76)	1.3376 (0.9440, 27.76)	1.9708 (1.6720, 8.96)	20.9941 (0.0039, 0.04)
LW	0.9260 (0.0030, 93.64)	1.6249 (0.0734, 23.40)	2.9017 (0.1812, 8.08)	26.8522 (0.0015, 0.08)
HSL	0.9339 (0.0332, 93.84)	1.2837 (0.8665, 29.40)	1.4359 (0.1949, 9.84)	12.2574 (0.0102, 0.04)
AM	0.9383 (0.0498, 93.76)	1.3376 (0.9440, 27.76)	1.9708 (1.5685, 8.96)	20.9941 (0.0039, 0.04)
GM	0.9392 (0.0565, 93.40)	1.2386 (2.7928, 30.20)	1.4474 (31.5700, 9.32)	05.0707 (0.0212, 0.04)
MED	0.9392 (0.0532, 93.60)	1.2723 (1.8752, 29.76)	1.4933 (31.5700, 9.52)	07.0164 (0.0203, 0.04)
KS	0.9336 (0.0309, 93.84)	1.4599 (0.3692, 25.68)	2.4969 (0.5450, 8.20)	33.5740 (0.0012, 0.08)
KS_AM	0.9358 (0.0309, 93.84)	1.4577 (0.2530, 26.08)	2.3629 (2.3302, 8.24)	25.4480 (0.7430, 0.04)
KS_MAX	0.9435 (0.0398, 93.92)	1.3101 (0.6745, 28.72)	1.6055 (0.6007, 9.16)	8.1732 (3.0535, 0.08)
KS_MED	0.9345 (0.0715, 93.84)	1.5204 (0.1668, 24.80)	3.0406 (2.0300, 7.84)	51.505 (0.1871, 0.04)
KS_GM	0.9348 (0.0345, 93.84)	1.5345 (0.1538, 24.72)	3.0023 (3.4393, 8.00)	48.3841 (0.0002, 0.04)
KM2	1.0903 (0.0359, 88.28)	1.1898 (1.9181, 32.64)	1.5825 (2.8401, 9.20)	14.5482 (0.0014, 0.04)
KM3	0.9840 (6.0050, 92.96)	1.1547 (58.5970, 32.12)	1.2792 (0.8296, 9.56)	3.0370 (0.0543, 0.04)
KM4	1.0865 (0.3674, 88.56)	1.2609 (0.8602, 29.72)	2.1760 (1.3919, 8.28)	38.6266 (0.0003, 0.04)
KM5	0.9740 (4.5079, 93.80)	1.2399 (1.4093, 30.68)	1.5905 (0.7200, 9.16)	09.6299 (0.0038, 0.08)
KM6	1.0868 (0.2322, 88.52)	1.2409 (1.0000, 30.24)	2.1364 (1.2528, 8.52)	37.4332 (0.0004, 0.04)
KM7	0.9737 (4.6621, 93.92)	1.2581 (1.1738, 30.36)	1.6190 (6.2880, 9.24)	10.9557 (0.0036, 0.08)
KM8	1.0492 (0.2254, 80.16)	1.1564 (4.6383, 34.44)	1.4168 (2.1108, 9.88)	10.2354 (0.0030, 0.04)
KM9	0.9473 (37.9604, 93.40)	1.1851 (2.3532, 32.00)	1.4116 (13.8060, 9.44)	7.7777 (0.0032, 0.08)
KM10	1.0730 (0.1355, 84.16)	1.2115 (2.3532, 31.20)	1.8736 (4.2749, 8.56)	25.5669 (0.0006, 0.04)
KM11	0.9391 (0.0549, 93.44)	1.3040 (1.3710, 28.84)	1.8124 (3.4567, 9.20)	16.4846 (0.0018, 0.08)
KM12	1.0709 (23.184, 83.84)	1.2097 (0.8781, 30.80)	2.0275 (1.2345, 8.48)	32.1960 (0.0005, 0.04)
GK	0.9433 (0.0705, 93.92)	1.4161 (1.4160, 25.72)	2.2847 (2.4356, 8.20)	27.8409 (1.0378, 0.08)
HMO	1.0005 (0.5086, 92.45)	1.1512 (0.4794, 26.54)	1.3997 (0.5679, 8.76)	9.2514 (8.0253, 0.04)
KD	0.9275 (0.0079, 91.98)	1.3581 (0.8946, 25.78)	2.0319 (0.9879, 8.92)	22.2854 (3.8341, 0.04)
CJH	1.0653 (11.7230, 92.34)	1.2061 (12.8190, 27.89)	1.5307 (13.5460, 8.76)	12.0446 (30.1750, 0.04)
FG	0.9777 (0.2594, 91.23)	1.1753 (2.6850, 27.92)	1.4181 (2.8790, 9.12)	8.9451 (5.4004, 0.08)

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table A4. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 50$, $p = 5$, and $\gamma = 0.7$. Condition number $\kappa = 28.37$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.7067	0.7181	0.9257	7.9302
HK	0.7069 (0.0240, 91.44)	0.7022 (0.5570, 8.12)	0.8447 (2.1865, 1.60)	4.1563 (44.5510, 1.24)
HKB	0.7071 (0.0501, 91.80)	0.6898 (1.2241, 9.64)	0.7865 (4.6033, 2.36)	3.0189 (28.4550, 1.32)
LW	0.7067 (0.0005, 90.80)	0.7178 (0.0108, 7.04)	0.9235 (0.0472, 1.48)	7.2028 (0.9094, 0.48)
HSL	0.7069 (0.0242, 91.44)	0.7016 (0.5657, 8.12)	0.8247 (2.5638, 1.48)	2.4236 (39.4080, 1.60)
AM	0.7071 (0.0501, 91.80)	0.6898 (1.2241, 9.64)	0.7865 (2.5638, 1.64)	3.0189 (28.4550, 1.32)
GM	0.7071 (0.0571, 91.88)	0.6806 (2.0736, 11.84)	0.7195 (4.6033, 2.36)	1.7582 (118.3500, 1.52)
MED	0.7072 (0.0641, 92.04)	0.6868 (1.6400, 10.96)	0.7675 (10.2230, 3.28)	1.9448 (106.1700, 1.64)
KS	0.7069 (0.0238, 91.44)	0.7036 (0.5006, 8.08)	0.8660 (7.0966, 3.08)	5.7339 (2.9518, 0.56)
KS_AM	0.7071 (0.0530, 91.88)	0.7066 (0.3863, 7.84)	0.8934 (1.4835, 1.56)	6.6141 (1.1676, 0.40)
KS_MAX	0.7073 (0.0796, 92.56)	0.7024 (0.5342, 8.08)	0.8549 (0.7058, 1.52)	4.6770 (4.1927, 0.56)
KS_MED	0.7071 (0.0492, 91.76)	0.7056 (0.4169, 7.84)	0.9030 (1.6563, 1.56)	7.2753 (0.5269, 0.36)
KS_GM	0.7071 (0.0487, 91.76)	0.7073 (0.3631, 7.80)	0.8991 (0.4985, 1.48)	7.2165 (0.5824, 0.36)
KM2	0.7959 (6.5729, 100.00)	0.6831 (1.3668, 9.56)	0.8911 (0.7070, 1.52)	7.1954 (0.4214, 0.36)
KM3	0.7091 (0.3012, 94.68)	0.6676 (7.2693, 15.12)	0.7073 (23.8640, 1.48)	2.2938 (137.6900, 1.40)
KM4	0.7636 (4.2624, 100.00)	0.6952 (0.7612, 8.16)	0.9061 (0.3795, 3.48)	7.7002 (0.1448, 0.36)
KM5	0.7086 (0.2375, 94.12)	0.6858 (0.3872, 9.76)	0.8209 (2.9563, 1.48)	3.9628 (8.9179, 0.92)
KM6	0.7603 (4.0412, 100.00)	0.6927 (0.8443, 8.28)	0.9021 (0.4517, 2.00)	7.6841 (0.1637, 0.36)
KM7	0.7087 (0.2513, 94.24)	0.6889 (1.2445, 9.76)	0.8376 (2.4653, 1.48)	4.0883 (8.0111, 0.92)
KM8	0.9818 (43.9325, 100.00)	0.6709 (2.0870, 11.20)	0.8897 (0.7168, 1.96)	7.1626 (0.4352, 0.36)
KM9	0.7074 (0.0901, 92.76)	0.6632 (3.1074, 13.52)	0.7776 (4.2204, 1.48)	4.5522 (4.4955, 0.56)
KM10	0.9030 (18.6343, 100.00)	0.6913 (0.8901, 8.24)	0.9045 (0.4149, 2.12)	7.5370 (0.2732, 0.36)
KM11	0.7071 (0.0563, 91.88)	0.6901 (1.1899, 9.56)	0.8317 (2.4992, 1.48)	5.0763 (3.7189, 0.56)
KM12	0.8911 (16.8071, 100.00)	0.6894 (0.9386, 8.28)	0.9036 (0.4220, 1.88)	7.5913 (0.2505, 0.36)
GK	0.7069 (0.0333, 91.60)	0.7020 (0.5652, 8.12)	0.8444 (2.1949, 1.48)	4.1521 (44.5610, 1.24)
HMO	0.7656 (4.3850, 100.00)	0.7364 (32.7499, 58.40)	0.5435 (81.3278, 1.60)	1.2348 (138.5700, 1.60)
KD	0.7069 (0.0301, 91.48)	0.6903 (1.2042, 9.64)	0.7870 (4.5830, 11.16)	3.0229 (28.4350, 1.32)
CJH	0.7144 (0.8160, 97.96)	0.6944 (70.5420, 45.00)	0.5912 (147.4000, 10.28)	1.9891 (114.2100, 1.48)
FG	0.7091 (0.2956, 94.60)	0.6517 (05.9959, 20.80)	0.6153 (18.3570, 4.60)	1.7616 (49.5940, 1.48)

Table A5. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 50$, $p = 5$, and $\gamma = 0.8$. Condition number $\kappa = 50.12$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.7754	0.8161	1.0648	9.4533
HK	0.7754 (0.0266, 60.32)	0.8087 (0.6000, 38.56)	1.01750 (2.1083, 23.56)	4.77240 (352.2000, 0.36)
HKB	0.7755 (0.0501, 61.40)	0.8076 (1.2116, 42.52)	0.98930 (4.4460, 26.12)	3.31000 (22.5420, 0.56)
LW	0.7754 (0.0005, 59.48)	0.8159 (0.0111, 34.12)	1.06300 (0.0484, 19.44)	8.50309 (0.9530, 0.20)
HSL	0.7754 (0.0267, 60.32)	0.8070 (0.6291, 38.68)	0.99420 (2.7644, 24.20)	2.49500 (32.8540, 0.48)
AM	0.7755 (0.0501, 61.40)	0.8076 (1.2116, 42.52)	0.98930 (4.4467, 26.12)	3.31000 (22.5490, 0.56)
GM	0.7755 (0.0548, 61.56)	0.8120 (1.9840, 46.04)	0.96060 (10.7210, 29.36)	1.66830 (91.4920, 0.76)
MED	0.7756 (0.0614, 61.72)	0.8095 (1.5500, 44.16)	0.97850 (7.0245, 27.20)	2.01050 (72.2960, 0.68)
KS	0.7754 (0.0265, 60.32)	0.8091 (0.5407, 38.2)	1.02619 (1.4740, 22.08)	6.83390 (2.5615, 0.20)
KS_AM	0.7755 (0.0470, 61.28)	0.8108 (0.3207, 36.40)	1.04240 (0.6254, 20.60)	8.05320 (0.9630, 0.16)
KS_MAX	0.7755 (0.0683, 62.12)	0.8077 (0.5641, 38.30)	1.01080 (1.8020, 22.48)	5.56520 (3.6691, 0.20)
KS_MED	0.7755 (0.0445, 61.12)	0.8110 (0.3127, 36.30)	1.05170 (0.3672, 19.96)	9.01690 (0.2741, 0.16)
KS_GM	0.7755 (0.0443, 61.16)	0.8114 (0.2841, 36.00)	1.04860 (0.4421, 20.16)	8.75150 (0.4528, 0.16)
KM2	0.8822 (6.2317, 100.00)	0.8017 (1.3228, 43.20)	1.03520 (0.7341, 20.56)	8.55220 (0.4577, 0.16)
KM3	0.7767 (0.2857, 70.48)	0.8252 (10.1870, 50.80)	0.95400 (46.5690, 29.72)	2.20640 (328.1800, 0.48)
KM4	0.8494 (4.3497, 100.00)	0.8033 (0.7904, 40.24)	1.04830 (0.3872, 20.00)	9.13830 (0.1613, 0.16)
KM5	0.7764 (0.2328, 68.64)	0.8064 (1.3477, 43.28)	0.99660 (2.9730, 23.92)	4.51400 (7.9870, 0.28)
KM6	0.8452 (4.1265, 100.00)	0.8029 (0.8755, 40.80)	1.04510 (0.4659, 20.24)	9.09630 (0.1874, 0.16)
KM7	0.7765 (0.2460, 69.08)	0.8061 (1.2070, 42.36)	1.00630 (2.4243, 23.36)	4.80110 (6.9678, 0.28)
KM8	1.0671 (39.4900, 100.00)	0.8008 (1.9541, 45.84)	1.03280 (0.7473, 20.56)	8.48080 (0.4980, 0.16)
KM9	0.7756 (0.0814, 62.88)	0.8153 (3.0747, 51.40)	0.97120 (4.5613, 26.60)	5.42800 (3.8988, 0.20)
KM10	1.0068 (19.378, 100.00)	0.8016 (0.9057, 40.96)	1.04730 (0.4091, 19.96)	8.89900 (0.3189, 0.16)
KM11	0.7755 (0.0542, 61.52)	0.8074 (1.1896, 42.04)	1.00220 (2.5720, 23.88)	6.11030 (3.1890, 0.20)
KM12	0.9953 (17.4860, 100.00)	0.8012 (0.9787, 41.20)	1.04680 (0.4198, 20.00)	8.95850 (0.2940, 0.16)
GK	0.7754 (0.0352, 60.68)	0.8086 (0.6075, 38.60)	1.01740 (2.1162, 23.56)	4.76620 (352.2900, 0.36)
HMO	0.8216 (2.9092, 99.20)	0.9433 (22.3590, 83.60)	0.96280 (56.9328, 45.6)	0.96700 (103.5200, 0.88)
KD	0.7755 (0.0301, 60.48)	0.8077 (1.1917, 42.42)	0.98960 (4.4268, 26.08)	3.31580 (22.5230, 0.56)
CJH	0.7894 (1.2570, 92.56)	0.9368 (59.4930, 84.31)	0.98740 (82.1500, 45.16)	1.86680 (762.3000, 0.68)
FG	0.7769 (0.2931, 70.40)	0.8302 (5.5836, 61.35)	0.94740 (16.3068, 36.6)	1.61010 (40.0510, 0.56)

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table A6. Simulated MSE, average ks and proportion of time (%) LSE perform better than ridge estimators for $n = 50$, $p = 5$, and $\gamma = 0.9$. Condition number $\kappa = 119.60$

Estimator	$\sigma = 0.1$	$\sigma = 0.5$	$\sigma = 1.0$	$\sigma = 5.0$
OLS	0.8678	1.0066	1.4255	18.4356
HK	0.8670 (0.0309, 27.44)	0.9735 (0.6019, 25.16)	1.2806 (1.8241, 15.44)	8.5312 (55.9952, 0.24)
HKB	0.8667 (0.0500, 28.60)	0.9590 (1.1721, 28.80)	1.1994 (3.9608, 18.08)	5.8085 (13.0289, 0.24)
LW	0.8678 (0.0006, 25.96)	1.0053 (0.0123, 20.36)	1.4168 (0.0537, 11.64)	14.7874 (1.0500, 0.12)
HSL	0.8670 (0.0309, 27.48)	0.9562 (0.7374, 25.52)	1.1796 (3.1439, 17.08)	2.7700 (35.1922, 0.24)
AM	0.8667 (0.0500, 28.60)	0.9590 (1.1721, 28.80)	1.1994 (3.9608, 18.08)	5.8085 (13.0289, 0.24)
GM	0.8667 (0.0531, 28.64)	0.9595 (2.2919, 33.08)	1.1402 (11.9880, 21.44)	2.2477 (57.2516, 0.28)
MED	0.8666 (0.0571, 28.92)	0.9583 (1.5774, 29.88)	1.1692 (7.5610, 19.28)	3.0584 (49.7520, 0.28)
KS	0.8670 (0.0307, 27.44)	0.9749 (0.5443, 24.72)	1.3011 (1.3187, 14.64)	11.1257 (2.2876, 0.16)
KS_AM	0.8669 (0.0388, 28.16)	0.9835 (0.2472, 22.08)	1.3471 (0.5417, 12.80)	13.4029 (0.9329, 0.08)
KS_MAX	0.8665 (0.0551, 28.72)	0.9593 (0.6567, 25.16)	1.2253 (2.0058, 15.52)	6.4475 (4.0782, 0.16)
KS_MED	0.8669 (0.0363, 27.88)	0.9899 (0.1721, 21.60)	1.3934 (0.2030, 11.88)	17.4920 (0.1406, 0.04)
KS_GM	0.8669 (0.0370, 27.92)	0.9894 (0.1787, 21.68)	1.3839 (0.2675, 12.16)	16.7151 (0.2668, 0.04)
KM2	1.0119 (5.7916, 99.96)	0.9348 (1.3595, 29.56)	1.2988 (0.8391, 13.28)	13.8647 (0.6424, 0.04)
KM3	0.8651 (0.2785, 40.32)	0.9691 (9.6830, 39.04)	1.1274 (164.0300, 22.64)	2.3765 (90.5992, 0.24)
KM4	0.9819 (4.4273, 99.80)	0.9475 (0.7989, 26.08)	1.3550 (0.4078, 12.48)	16.8254 (0.2065, 0.04)
KM5	0.8650 (0.2289, 38.00)	0.9503 (1.3896, 30.36)	1.2126 (2.9486, 16.92)	6.2501 (6.2615, 0.20)
KM6	0.9784 (4.2866, 99.64)	0.9445 (0.9231, 27.04)	1.3455 (0.4886, 12.72)	16.5862 (0.2421, 0.04)
KM7	0.8650 (0.2370, 38.52)	0.9516 (1.1690, 28.60)	1.2323 (2.3762, 15.92)	6.8584 (5.6013, 0.16)
KM8	1.1763 (34.1253, 100.00)	0.9312 (2.0914, 32.76)	1.2820 (0.9396, 13.28)	13.2874 (0.7194, 0.04)
KM9	0.8664 (0.0780, 29.76)	0.9579 (3.4085, 39.32)	1.1441 (4.8109, 19.88)	6.2412 (4.2808, 0.16)
KM10	1.1438 (20.0632, 100.00)	0.9409 (0.9377, 26.92)	1.3513 (0.4309, 12.52)	15.9926 (0.3402, 0.04)
KM11	0.8667 (0.0525, 28.64)	0.9557 (1.1940, 29.08)	1.2215 (2.5096, 16.44)	8.6147 (3.0609, 0.16)
KM12	1.1380 (18.8572, 100.00)	0.9392 (1.0620, 27.84)	1.3524 (0.4330, 12.60)	16.3197 (0.3018, 0.04)
GK	0.8669 (0.0389, 27.96)	0.9731 (0.6089, 25.20)	1.2800 (1.8316, 15.44)	8.5085 (56.0051, 0.24)
HMO	0.8936 (1.4404, 82.56)	1.0294 (12.1690, 58.72)	1.0577 (29.8000, 27.80)	1.5332 (42.5577, 0.28)
KD	0.8672 (0.0300, 27.56)	0.9597 (1.1522, 28.72)	1.2004 (3.9409, 18.00)	5.8310 (13.0090, 0.24)
CJH	0.9395 (3.2631, 94.16)	1.0615 (30.2460, 65.08)	1.0986 (64.1490, 28.36)	3.0802 (840.5210, 0.24)
FG	0.8653 (0.2857, 41.08)	0.9642 (4.7476, 45.12)	1.0884 (12.3090, 24.48)	2.5102 (21.1200, 0.24)

Table A7. Simulated MSE, average ks and proportion of time (%) LSE perform better than proposed new ridge estimators for different values of n , $p = 5$, $\sigma = 3.0$, and $\gamma = 0.9$

Estimator	$n = 10$	$n = 20$	$n = 50$
OLS	67.1158	22.6167	7.7726
HK	26.2573 (78.0160, 0.12)	10.0504 (83.0563, 0.32)	3.8897 (20.1510, 1.08)
HKB	20.9523 (1.7574, 0.12)	7.1599 (3.9477, 0.36)	2.8018 (10.1280, 1.52)
LW	15.9023 (1.3047, 0.12)	11.5401 (1.0943, 0.28)	6.5964 (0.4158, 0.36)
HSL	9.6370 (6.0834, 0.16)	4.0571 (10.3777, 0.32)	1.4353 (25.3280, 1.64)
AM	20.9523 (1.7574, 0.12)	7.1599 (3.9477, 0.36)	2.8018 (10.1280, 1.52)
GM	5.0207 (9.8421, 0.16)	2.8726 (17.2089, 0.28)	1.6304 (36.1640, 1.92)
MED	6.0914 (12.5365, 0.16)	3.6802 (16.7250, 0.36)	1.7539 (30.9620, 1.52)
KS	26.9777 (1.1632, 0.12)	11.4083 (1.2309, 0.20)	4.5503 (2.3287, 0.56)
KS_AM	8.3030 (1.0881, 0.12)	10.4941 (0.7219, 0.08)	5.0463 (1.0467, 0.32)
KS_MAX	2.6825 (4.3804, 0.12)	3.8964 (2.9958, 0.20)	2.6053 (4.5236, 0.56)
KS_MED	20.9093 (0.3330, 0.12)	19.1547 (0.1196, 0.04)	7.0904 (0.1791, 0.28)
KS_GM	22.9548 (0.3001, 0.12)	16.7338 (0.2319, 0.04)	6.7395 (0.2876, 0.32)
KM2	4.0575 (2.5361, 0.12)	6.7336 (1.2331, 0.04)	5.2090 (0.7023, 0.28)
KM3	1.9326 (49.8389, 0.12)	02.1282 (61.7303, 0.36)	1.7384 (59.9134, 1.52)
KM4	13.7695 (0.5735, 0.12)	13.7225 (0.3724, 0.04)	6.7678 (0.2385, 0.28)
KM5	5.7004 (2.5166, 0.12)	4.6168 (3.4765, 0.32)	2.8852 (5.1752, 0.84)
KM6	13.7148 (0.6410, 0.12)	13.1157 (0.4334, 0.04)	6.7395 (0.2611, 0.28)
KM7	6.2518 (2.4318, 0.12)	5.1636 (3.2115, 0.32)	2.9637 (4.6849, 0.80)
KM8	2.6439 (10.0721, 0.12)	5.0384 (2.1737, 0.04)	4.9405 (0.7863, 0.32)
KM9	2.2439 (5.1076, 0.12)	3.5706 (3.3286, 0.24)	2.4176 (5.2458, 0.68)
KM10	9.5348 (0.9021, 0.12)	10.6497 (0.6062, 0.04)	6.5327 (0.3014, 0.28)
KM11	8.7200 (1.6719, 0.12)	6.7272 (1.8406, 0.20)	3.2644 (3.4917, 0.60)
KM12	14.0551 (0.7741, 0.12)	12.0833 (0.5238, 0.04)	6.7666 (0.2596, 0.28)
GK	20.2413 (78.082, 0.12)	9.6857 (83.0930, 0.32)	3.8801 (20.1590, 1.08)
HMO	11.5302 (3.0492, 0.16)	3.1834 (8.8390, 0.32)	1.3795 (37.4880, 2.28)
KD	28.8528 (1.6649, 0.12)	7.4149 (3.8977, 0.36)	2.8119 (10.1080, 1.52)
CJH	13.6394 (3.8924, 0.16)	4.2477 (13.8341, 0.28)	1.8765 (38.3210, 1.96)
FG	9.0114 (2.2648, 0.12)	3.4396 (6.1783, 0.36)	1.6613 (18.5430, 1.64)
KB1	3.6922 (13.5707, 0.16)	2.5210 (20.7526, 0.36)	1.4909 (36.8980, 1.96)
KB2	6.4905 (4.7393, 0.16)	3.0329 (9.6154, 0.36)	1.6045 (23.6380, 1.84)
KB3	1.8130 (57.2684, 0.16)	1.8326 (76.6178, 0.28)	1.2435 (107.5580, 2.32)
KB4	5.6967 (4.5998, 0.12)	2.8649 (9.7842, 0.36)	1.5804 (24.0460, 1.84)
KB5	8.3678 (3.2453, 0.12)	3.2945 (7.7060, 0.36)	1.6849 (19.6580, 1.72)

SOME RIDGE REGRESSION ESTIMATORS AND THEIR PERFORMANCES

Table A8. Simulated MSE, average ks and proportion of time (%) LSE perform better than proposed new ridge estimators for different values of n , $p = 5$, $\sigma = 10.0$, and $\gamma = 0.9$

Estimator	$n = 10$	$n = 20$	$n = 50$
OLS	737.0407	242.9571	78.1037
HK	272.0000 (1.3987, 0)	93.0000 (8.6474, 0)	29.5000 (4.1928, 0.00)
HKB	223.1200 (1.8062, 0)	67.8198 (4.5231, 0)	20.9440 (13.9677, 0.04)
LW	153.6900 (1.6560, 0)	103.8625 (1.6510, 0)	53.5800 (1.3756, 0.00)
HSL	116.1700 (6.1421, 0)	50.8556 (11.0230, 0)	11.7460 (48.0250, 0.04)
AM	223.1200 (1.8062, 0)	67.8198 (4.5230, 0)	20.9440 (13.9670, 0.04)
GM	41.7100 (11.4337, 0)	15.4604 (26.2730, 0)	5.2872 (76.8460, 0.00)
MED	58.2900 (12.5456, 0)	28.2886 (24.1990, 0)	8.7540 (69.4340, 0.04)
KS	279.7900 (1.3952, 0)	109.0829 (1.4290, 0)	38.4023 (2.8360, 0.00)
KS_AM	74.7100 (1.1902, 0)	101.9246 (0.7850, 0)	45.2139 (1.2170, 0.00)
KS_MAX	17.9800 (4.8874, 0)	28.5829 (3.3113, 0)	17.8669 (5.3720, 0.00)
KS_MED	223.0900 (0.3343, 0)	204.5300 (0.1190, 0)	70.6056 (0.1790, 0.00)
KS_GM	240.9700 (0.3085, 0)	176.4906 (0.2370, 0)	66.4137 (0.2980, 0.00)
KM2	37.2600 (2.5292, 0)	65.9034 (1.2148, 0)	50.4013 (0.6740, 0.00)
KM3	10.9600 (56.5641, 0)	7.9644 (70.3667, 0)	5.2666 (192.4500, 0.04)
KM4	153.2700 (0.5298, 0)	153.0248 (0.3210, 0)	69.6911 (0.1750, 0.00)
KM5	51.4400 (2.7141, 0)	35.3579 (4.1621, 0)	18.9584 (7.3710, 0.00)
KM6	150.7500 (0.6215, 0)	140.6503 (0.4110, 0)	68.2656 (0.2180, 0.00)
KM7	60.2200 (2.5963, 0)	45.2184 (3.6642, 0)	21.9880 (6.5460, 0.00)
KM8	21.5800 (10.0380, 0)	46.3192 (2.1200, 0)	47.6514 (0.7460, 0.00)
KM9	15.7200 (5.3348, 0)	27.3427 (3.4367, 0)	17.5897 (5.4840, 0.00)
KM10	100.3600 (0.8550, 0)	111.8558 (0.5840, 0)	65.2522 (0.2820, 0.00)
KM11	85.6400 (1.7224, 0)	63.1383 (1.8982, 0)	26.9247 (3.7130, 0.00)
KM12	153.5700 (0.7523, 0)	127.6506 (0.5070, 0)	67.7397 (0.2410, 0.00)
GK	208.0000 (1.3987, 0)	90.0000 (8.6474, 0)	29.5000 (4.1920, 0.00)
HMO	119.9000 (2.9913, 0)	23.8446 (9.4118, 0)	5.5032 (42.1730, 0.00)
KD	309.9300 (1.7137, 0)	70.5684 (4.4732, 0)	21.0470 (13.9400, 0.04)
CJH	143.4400 (3.7576, 0)	35.9131 (13.0910, 0)	11.3326 (37.6000, 0.00)
FG	91.7100 (2.2766, 0)	26.5555 (6.4474, 0)	8.7287 (20.7300, 0.04)
KB1	28.3800 (14.9280, 0)	12.9418 (24.9650, 0)	5.0918 (73.2000, 0.04)
KB2	61.270 (4.8625, 0)	20.2705 (11.6430, 0)	6.3075 (36.1800, 0.04)
KB3	9.9300 (63.2258, 0)	6.2206 (88.2720, 0)	2.9234 (257.3900, 0.00)
KB4	51.3900 (5.0065, 0)	17.4754 (11.9870, 0)	6.1536 (35.6300, 0.04)
KB5	83.2800 (3.3955, 0)	23.2968 (8.7820, 0)	7.5708 (26.6600, 0.04)

Construction of Efficiency-Balanced Design Using Factorial Design

Rajarathinam Arunachalam
Manonmaniam Sundaranar University
Tirunelveli, India

Mahalakshmi Sivasubramanian
Manonmaniam Sundaranar University
Tirunelveli, India

Dilip Kumar Ghosh
Saurashtra University
Rajkot, India

Two different methods are proposed for the construction of an efficiency balanced design. Method 1 discusses the construction of efficiency balanced design by deleting the control treatment and method 2 discusses the construction of efficiency balanced design by deleting the control treatment as well as all the main effect treatment combinations of a $2n$ symmetrical factorial experiment. Numerical examples are given.

Keywords: Block designs, C-matrix, M-matrix, efficiency factor, factorial design

---

Introduction

The concept of efficiency balance is due to Jones (1959) and the nomenclature “efficiency balance” is due to Puri and Nigam (1975) and Williams (1975). Block designs with v treatments and b blocks are considered. It is assumed that the i^{th} treatment is replicated r_i times, $i = 1, 2, 3, \dots, v$ and the j^{th} block contains k_j (not necessarily distinct), treatments, $j = 1, 2, 3, \dots, b$. Let $\mathbf{r} = [r_1, r_2, r_3, \dots, r_v]'$, $\mathbf{k} = [k_1, k_2, k_3, \dots, k_b]'$, $R = \text{diag}(r_1, r_2, r_3, \dots, r_v)$, $K = \text{diag}(k_1, k_2, k_3, \dots, k_b)$, and $\mathbf{N}$ be the $v \times b$ incidence matrix of the design. If $\mathbf{T}$ denotes the column vector of treatment totals then $\mathbf{s}'\mathbf{T}$ is called a contrast of treatment totals if $\mathbf{s}'\mathbf{r} = 0$, where $\mathbf{s}$ is a column vector. The intra-block component of $\mathbf{s}'\mathbf{T}$ is defined by Jones (1959) as $\mathbf{s}'\mathbf{Q}$ where $\mathbf{Q}$ is the vector of adjusted treatment totals, given by $\mathbf{Q} = \mathbf{T} - \mathbf{N}\mathbf{K}^{-1}\mathbf{B}$, $\mathbf{B}$ being the vector of block totals.

Jones (1959) showed that if $\mathbf{s}$ is a right eigenvector of the matrix $\mathbf{M} = \mathbf{R}^{-1}\mathbf{N}\mathbf{K}^{-1}\mathbf{N}'$ corresponding to an eigenvalue $\mu (\neq 1)$, then the loss of information on the ‘intra-block component’ of $\mathbf{s}'\mathbf{T}$ is μ so that the efficiency-factor of the ‘intra-block component’ is $1 - \mu$.

---

*Dr. Rajarathinam is a Professor of Statistics. Email at arrathinam@yahoo.com.
Mahalakshmi Sivasubramanian is a Part-Time Research Scholar. Email at seethamaha06@gmail.com.
Dilip Kumar Ghosh is a UGC BSR Fellow. Email at ghosh_dkg@yahoo.com.*

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

Because $\mathbf{s}'\mathbf{Q}$ (the intra-block component of $\mathbf{s}'\mathbf{T}$) is a function of observations and not of parameters (treatment effects) the concept of 'loss of information' or 'efficiency-factor' of $\mathbf{s}'\mathbf{Q}$ is a little confusing when viewed from the classical definition of loss of information, referring to the loss incurred in estimating a certain contrast of treatment effects through a design, in relation to an orthogonal design.

A block design for which every contrast has the same loss of information (or, equivalently, same efficiency-factor) may be termed Efficiency Balanced. The concept of efficiency balance is different from the one used commonly, according to which design is balanced if every elementary contrast is estimated through the design with the same variance. To avoid confusion, the latter concept is called Variance-Balance (see e.g., [Hedayat and Federer, 1974](#)).

Calinski (1971) and Puri and Nigam (1975) established a sufficient condition for a design to be efficiency balanced is that its $\mathbf{M}$ matrix, given by

$$\mathbf{M} = \mu\mathbf{I} + (1 - \mu)\mathbf{1}\mathbf{r}' / n \quad (1)$$

where n is the total number of observations in the design. That (1) is necessary as well for a design to be efficiency balanced was shown by Williams (1975). He also showed that, for more than two varieties, an efficiency balanced design was also a variance balanced design if and only if the design is equi-replicated. Puri and Nigam (1975) gave a note on efficiency balanced design. Dey et al. (1981) proved that a necessary and sufficient condition for a design to be efficiency balanced is that (1) holds.

Mukerjee and Saha (1990) derived some optimality results on efficiency balanced designs. Gupta and Prasad (1991) gave a method for constructing general efficiency balanced designs with equal and unequal block sizes. Gupta (1992) gave a method for constructing efficiency balanced designs through BIB and GD designs.

Ceranka and Graczyk (2009) discussed some problems for a class of EB-BD based on balanced incomplete block designs with repeated blocks. Awad and Banerjee (2012) gave a method for constructing variance and efficiency balanced block designs with repeated blocks which are based on the incidence matrices of the known balanced incomplete block designs with repeated blocks. Sun and Tang (2010) gave the optimal efficiency balanced designs and their constructions.

Purpose of the Study

Efficiency balanced design using 2^n symmetrical factorial design by deleting control treatment

The construction of unequal block sizes and equi-replicated binary EB designs from symmetrical factorial designs are discussed. First, consider the following lemma without proof.

Lemma 1 In a 2^n symmetrical factorial experiment, delete the control treatment. For an example, let $n = 3$. The $2^3 = 8$ treatment combinations are

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Delete a treatment combination whose level of all factor is zero. That is, delete a control treatment. Keep the remaining treatment combinations as such. Finally, the treatment combinations are

0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Let the matrix $\mathbf{N}$ be the combinations of $2^n - 1$ treatment combinations. Finally the matrix (transpose of $\mathbf{N}$) becomes the incidence matrix of efficiency balanced design. Construction of an efficiency balanced design is shown in the theorem that follows.

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

Theorem If there exists a 2^n symmetrical factorial experiment then there always exists an unequal block sizes, equi-replicated, binary EB design, by deleting the control treatment with the following parameters

$$v = n, b = 2^n - 1, r = 2^{n-1}, \mathbf{k} = \left(\underbrace{1, \dots, 1}_{{}^nC_1 \text{ times}}, \underbrace{2, \dots, 2}_{{}^nC_2 \text{ times}}, \underbrace{3, \dots, 3}_{{}^nC_3 \text{ times}}, \dots, \underbrace{v-1, \dots, v-1}_{{}^nC_{n-1} \text{ times}}, \underbrace{v}_{{}^nC_n \text{ times}} \right)$$

Proof Consider a 2^n symmetrical factorial experiment. This has 2^n treatment combinations. Considering n factors as rows and 2^n treatment combinations as columns, and then using the Lemma 1, we have the following incidence matrix of a design d. The incidence matrix $\mathbf{N}$ is given as

$$\mathbf{N} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 1 \\ 1 & 0 & 1 & \dots & \dots & 1 \\ 1 & 0 & 1 & \dots & \dots & 1 \\ 0 & 1 & 0 & \dots & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & 1 & 1 & \dots & \dots & 1 \\ 1 & 0 & 0 & \dots & \dots & 1 \end{bmatrix} \quad (2)$$

Since we have n rows and considering these as treatments, obviously $v = n$.

For the incidence matrix $\mathbf{N}$, among ${}^nC_1$ columns, in each column the element 1 will occur once and 0 will occur $(n - 1)$ times. Similarly, for ${}^nC_2$ columns, the element 1 will occur two times and the element 0 will occur $(n - 2)$ times in each column, and so on. Moreover, there will be one column whose elements are all unity. Hence the number of blocks is

$$b = {}^nC_1 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n$$

Because ${}^nC_0 + {}^nC_1 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n = 2^n$, $b = 2^n - 1$.

Among $(2^n - 1)$ columns, ${}^nC_1$ columns have block size 1, ${}^nC_2$ columns have block size 2, and so on.

Hence

$$\mathbf{k} = \left(\underbrace{1, \dots, 1}_{{}^nC_1 \text{ times}}, \underbrace{2, \dots, 2}_{{}^nC_2 \text{ times}}, \underbrace{3, \dots, 3}_{{}^nC_3 \text{ times}}, \dots, \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1} \text{ times}}, \underbrace{n}_{{}^nC_n \text{ times}} \right)$$

Factors having level 0 occur $(2^{n-1} - 1)$ times, and factors having level 1 occur 2^{n-1} times. Since we have considered rows as treatments, there are $v = n$ treatments. Similarly, we have considered columns as blocks and hence we have $b = (2^n - 1)$ blocks. In each row, one occurs 2^{n-1} times, so the number of replications is $r = 2^{n-1}$.

Using the incidence matrix $\mathbf{N}$ shown in (2), we have the following C-matrix.

$$\mathbf{C} = \begin{pmatrix} \alpha & \beta & \beta & \beta & \cdots & \beta \\ \beta & \alpha & \beta & \beta & \cdots & \beta \\ \beta & \beta & \alpha & \beta & \cdots & \beta \\ \beta & \beta & \beta & \alpha & \cdots & \beta \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \beta & \beta & \cdots & \alpha \end{pmatrix}_{(v \times v)} \quad (3)$$

where

$$\alpha = r_i - \sum_j \left[\frac{n_{ij}^2}{k_j} \right], \beta = - \sum_j \left[\frac{n_{ij} n_{i|j}}{k_j} \right]$$

The eigenvalue of the C-matrix is

$$\theta = \alpha \left[\frac{v}{v-1} \right]$$

with multiplicities $(v - 1)$, where v is the number of treatments. Also, $\mathbf{M} = \mathbf{I} - \mathbf{C}R^{-1}$.

After some simplifications, the matrix $\mathbf{M}$ is obtained as

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

$$\mathbf{M} = \begin{pmatrix} 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & \dots & -\frac{\beta}{r} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & 1 - \frac{\alpha}{r} \end{pmatrix}_{(v \times v)} \quad (4)$$

Thus $\mathbf{MJ} = \mathbf{J}$, where $\mathbf{J}$ is the unit vector of order $(v \times 1)$, and the design having $\mathbf{N}$ as the incidence matrix is the efficiency balanced design.

The incidence matrix $\mathbf{N}$ in (2) gives unequal block sizes, equi-replicated and binary EB designs with parameters

$$v = n, b = 2^n - 1, r = 2^{n-1}, \mathbf{k} = \left(\underbrace{1, \dots, 1}_{{}^nC_1 \text{ times}}, \underbrace{2, \dots, 2}_{{}^nC_2 \text{ times}}, \underbrace{3, \dots, 3}_{{}^nC_3 \text{ times}}, \dots, \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1} \text{ times}}, \underbrace{n}_{{}^nC_n \text{ times}} \right)$$

Calculation of efficiency factor

The M-matrix of the efficiency balanced design is

$$\mathbf{M} = \mu \mathbf{I} + (1 - \mu) \mathbf{J} \mathbf{r}' / n$$

where μ is the loss of information, $\mathbf{I}$ is the identity matrix of order $(v \times v)$, $\mathbf{J}$ is the unit vector of order $(v \times 1)$, $\mathbf{r}'$ is the row vector of order $(1 \times v)$, and n is the number of observations.

$$\mathbf{M} = \begin{bmatrix} \mu & 0 & 0 & \dots & 0 \\ 0 & \mu & 0 & \dots & 0 \\ 0 & 0 & \mu & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \mu \end{bmatrix} + \left\{ (1 - \mu) \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} r & r & r & \dots & r \end{bmatrix} \right\} \bigg/ \sum_{i=1}^v r_i$$

$$\begin{aligned}
 \mathbf{M} = & \mu \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \end{bmatrix} \\
 & - \mu \begin{bmatrix} r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r/\sum r_i & r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \end{bmatrix} \\
 \mathbf{M} = & \mu \begin{bmatrix} 1-r/\sum r_i & 1-r/\sum r_i & \cdots & 1-r/\sum r_i \\ 1-r/\sum r_i & 1-r/\sum r_i & \cdots & 1-r/\sum r_i \\ \vdots & \vdots & \ddots & \vdots \\ 1-r/\sum r_i & 1-r/\sum r_i & \cdots & 1-r/\sum r_i \end{bmatrix} \\
 & + \begin{bmatrix} r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \\ \vdots & \vdots & \ddots & \vdots \\ r/\sum r_i & r/\sum r_i & \cdots & r/\sum r_i \end{bmatrix}
 \end{aligned} \tag{5}$$

Equating (4) and (5),

$$\mu = \left(\sum_{i=1}^v r_i - \frac{\alpha \sum_{i=1}^v r_i}{r} - r \right) / \left(\sum_{i=1}^v r_i - r \right)$$

The efficiency factor is $E = 1 - \mu$. By putting the value of μ , and after some simplifications, the efficiency factor E can be written as

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

$$E = \frac{\alpha \sum_{i=1}^v r_i / r}{\sum_{i=1}^v r_i - r}$$

Numerical Example

In a 2^4 factorial design, the incidence matrix after deleting the control treatment is given by

$$\mathbf{N} = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Here $v = n = 4$. The above incidence matrix is of the order (4×15) which gives $\mathbf{R}$ and $\mathbf{K}$ matrices of order (4×4) and (15×15) , respectively, where

$$\mathbf{R} = \text{diag}(8 \quad 8 \quad 8 \quad 8)$$

$$\mathbf{K} = \text{diag} \left(\underbrace{1 \quad 1 \quad 1}_{{}^4C_1 \text{ times}} \quad \underbrace{2 \quad 2 \quad 2 \quad 2}_{{}^4C_2 \text{ times}} \quad \underbrace{23 \quad 3 \quad 3 \quad 3}_{{}^4C_3 \text{ times}} \quad \underbrace{4}_{{}^4C_4 \text{ times}} \right)$$

The C-matrix is given by

$$\mathbf{C} = \begin{bmatrix} 17/4 & -17/12 & -17/12 & -17/12 \\ -17/12 & 17/4 & -17/12 & -17/12 \\ -17/12 & -17/12 & 17/4 & -17/12 \\ -17/12 & -17/12 & -17/12 & 17/4 \end{bmatrix}$$

The non-zero eigenvalue of the $\mathbf{C}$ matrix is $\theta = 17/3$.

The $\mathbf{M}$ -matrix of the efficiency balanced design can be obtained by substituting all the values of $\mathbf{N}$, $\mathbf{R}$, and $\mathbf{K}$ in $\mathbf{M} = \mathbf{R}^{-1}\mathbf{N}\mathbf{K}^{-1}\mathbf{N}'$ and, after some simplifications, the matrix $\mathbf{M}$ of the required design is

$$\mathbf{M} = \begin{bmatrix} 15/32 & 17/96 & 17/96 & 17/96 \\ 17/96 & 15/32 & 17/96 & 17/96 \\ 17/96 & 17/96 & 15/32 & 17/96 \\ 17/96 & 17/96 & 17/96 & 15/32 \end{bmatrix} \quad (6)$$

Obviously this matrix satisfies the conditions of efficiency balanced design i.e. $\mathbf{MJ} = \mathbf{J}$, where $\mathbf{J}$ is the $(v \times 1)$ unit vector.

The efficiency factor is calculated by using the formula

$$\mathbf{M} = \mu \mathbf{I} + (1 - \mu) \mathbf{J} \mathbf{r}' / n$$

$$\mathbf{M} = \begin{bmatrix} \mu & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & \mu \end{bmatrix} + \left\{ (1 - \mu) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 8 & 8 & 8 & 8 \end{bmatrix} \right\} / 32$$

$$\mathbf{M} = \begin{bmatrix} \mu & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & \mu \end{bmatrix} + \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} - \mu \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \quad (7)$$

Equating (6) and (7), we get $\mu = 7/24$.

The design with the above incidence matrix gives efficiency balanced design with parameters

$$v = 4, b = 15, \mathbf{r} = (8 \ 8 \ 8 \ 8),$$

$$\mathbf{k} = \left(\underbrace{1 \ 1 \ 1 \ 12}_{{}^4C_1 \text{ times}} \underbrace{2 \ 2 \ 2 \ 2 \ 23}_{{}^4C_2 \text{ times}} \underbrace{3 \ 3 \ 3}_{{}^4C_3 \text{ times}} \underbrace{4}_{{}^4C_4 \text{ times}} \right)$$

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

having efficiency factor $E = (1 - \mu) = 17/24$.

Efficiency balanced design using 2^n symmetrical factorial design by deleting control and all main effect treatments

Lemma 2 Consider a 2^n symmetrical factorial experiment. From the 2^n treatment combinations, delete the control treatments as well as all main effects. Hence we have $2^n - n - 1$ treatment combinations as the blocks in the required design.

As an example, let $n = 3$. The $2^3 = 8$ treatment combinations are

0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Delete any treatment combinations where the levels of all factors is zero. That is, delete any control treatments. Next, delete all treatment combinations where the level of one factor is one while the levels of all other factors are zero. That is, delete the main effects. Keep the remaining treatment combinations as such. Finally, the treatment combinations are

0	1	1
1	0	1
1	1	0
1	1	1

Let the matrix $\mathbf{N}$ be the combinations of $2^n - n - 1$ treatment combinations. Finally, the matrix (transpose of $\mathbf{N}$) becomes the incidence matrix of an efficiency balanced design. Construction of an efficiency balanced design is given in the theorem that follows.

Theorem If there exists 2^n symmetrical factorial experiments then there always exists an unequal block sizes, equi-replicated, binary EB design by deleting the control treatment and main effects with the following parameters:

$$v = n, b = 2^n - n - 1, r = 2^{n-1} - 1, \text{ and}$$

$$\mathbf{k} = \left(\underbrace{2, \dots, 2}_{{}^nC_2 \text{ times}}, \underbrace{3, \dots, 3}_{{}^nC_3 \text{ times}}, \dots, \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1} \text{ times}}, \underbrace{n}_{{}^nC_n \text{ times}} \right)$$

Proof Consider a 2^n symmetrical factorial experiment. This has 2^n treatment combinations. Considering n factors as rows and 2^n treatment combinations as columns and then using the [Lemma 2](#) we have the following incidence matrix of a design d. The incidence matrix $\mathbf{N}$ is given as

$$\mathbf{N} = \begin{bmatrix} 0 & 1 & 0 & \dots & \dots & 1 \\ 1 & 0 & 1 & \dots & \dots & 1 \\ 1 & 0 & 1 & \dots & \dots & 1 \\ 0 & 1 & 0 & \dots & \dots & 1 \\ \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & 1 & 1 & \dots & \dots & 1 \\ 1 & 0 & 0 & \dots & \dots & 1 \end{bmatrix} \quad (8)$$

Since we have n rows and, considering these as treatments, $v = n$.

For the incidence matrix $\mathbf{N}$, among ${}^nC_2$ columns, the element 1 will occur two times and the element 0 will occur $(n-2)$ times. Similarly for ${}^nC_3$ columns, the element 1 will occur three times and the element 0 will occur $(n-3)$ times in each column, and so on. Moreover there will be one column whose all elements are unity. Hence the number of blocks is

$$b = {}^nC_2 + {}^nC_3 + \dots + {}^nC_n$$

Because ${}^nC_0 + {}^nC_1 + {}^nC_2 + {}^nC_3 + \dots + {}^nC_n = 2^n$, $b = 2^n - 1$.

Among $2^n - n - 1$ columns, ${}^nC_2$ columns have block size 2 while ${}^nC_3$ columns have block size 3, and so on.

Hence

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

$$\mathbf{k} = \left(\underbrace{2, \dots, 2}_{{}^nC_2 \text{ times}}, \underbrace{3, \dots, 3}_{{}^nC_3 \text{ times}}, \dots, \underbrace{n-1, \dots, n-1}_{{}^nC_{n-1} \text{ times}}, \underbrace{n}_{{}^nC_n \text{ times}} \right)$$

Factors having level 0 occur n times, while factors having level 1 occur $2^{n-1} - 1$ times. Because rows are considered as treatments, there are $v = n$ treatments. Similarly, columns are considered as blocks, so there are $b = 2^n - n - 1$ blocks. In each row, 1 occurs $2^{n-1} - 1$ times, so the number of replication is $r = 2^{n-1} - 1$.

Using the incidence matrix $\mathbf{N}$ shown in (8), we have the following C-matrix.

$$\mathbf{C} = \begin{pmatrix} \alpha & \beta & \beta & \beta & \cdots & \beta \\ \beta & \alpha & \beta & \beta & \cdots & \beta \\ \beta & \beta & \alpha & \beta & \cdots & \beta \\ \beta & \beta & \beta & \alpha & \cdots & \beta \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \beta & \beta & \cdots & \alpha \end{pmatrix}_{(v \times v)} \quad (9)$$

where

$$\alpha = r_i - \sum_j \left[\frac{n_{ij}^2}{k_j} \right], \beta = - \sum_j \left[\frac{n_{ij} n_{ilj}}{k_j} \right]$$

The eigenvalue of the $\mathbf{C}$ matrix in (9) is

$$\theta = \alpha \left[\frac{v}{v-1} \right]$$

with multiplicities $(v - 1)$, where v is the number of treatments. Also, $\mathbf{M} = \mathbf{I} - \mathbf{C}\mathbf{R}^{-1}$.

After some simplifications, the matrix $\mathbf{M}$ is obtained as

$$\mathbf{M} = \begin{pmatrix} 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & -\frac{\beta}{r} & \dots & -\frac{\beta}{r} \\ -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & 1 - \frac{\alpha}{r} & \dots & -\frac{\beta}{r} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & -\frac{\beta}{r} & \dots & 1 - \frac{\alpha}{r} \end{pmatrix}_{(v \times v)} \quad (10)$$

It can be verified that $\mathbf{MJ} = \mathbf{J}$ where $\mathbf{J}$ is the unit vector of order $(v \times 1)$ and hence the design having $\mathbf{N}$ as the incidence matrix is the efficiency balanced design.

This indicates the incidence matrix $\mathbf{N}$ in (8) gives unequal block sizes, equi-replicated and binary EB designs with parameters

$$v = n, b = 2^n - n - 1, r = 2^{n-1} - 1, \mathbf{k} = \left(\underbrace{2, \dots, 2}_{n C_2 \text{ times}}, \underbrace{3, \dots, 3}_{n C_3 \text{ times}}, \underbrace{v-1, \dots, v-1}_{n C_{n-1} \text{ times}}, \underbrace{v}_{n C_n \text{ times}} \right)$$

Numerical Example

In a 2^4 factorial design, the incidence matrix after deleting the control treatment and main effects is given as

$$\mathbf{N} = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Here $v = n = 4$. The above incidence matrix is of the order (4×11) which gives R and K matrices of order (4×4) and (11×11) , respectively, where

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

$$\mathbf{R} = \text{diag}(7 \ 7 \ 7 \ 7)$$

$$\mathbf{K} = \text{diag} \left(\underbrace{2 \ 2 \ 2 \ 2 \ 2}_{{}^4C_2 \text{ times}} \underbrace{23 \ 3 \ 3 \ 3}_{{}^4C_3 \text{ times}} \underbrace{4}_{{}^4C_4 \text{ times}} \right)$$

The C-matrix is given by

$$\mathbf{C} = \begin{bmatrix} 17/4 & -17/12 & -17/12 & -17/12 \\ -17/12 & 17/4 & -17/12 & -17/12 \\ -17/12 & -17/12 & 17/4 & -17/12 \\ -17/12 & -17/12 & -17/12 & 17/4 \end{bmatrix}$$

The non-zero eigenvalue of the C-matrix is $\theta = 17/3$.

The matrix $\mathbf{M}$ of the efficiency balanced design can be obtained by substituting all the values of $\mathbf{N}$, R , and K in $\mathbf{M} = R^{-1}\mathbf{N}K^{-1}\mathbf{N}'$ and, after some simplifications, the M-matrix of the required design is

$$\mathbf{M} = \begin{bmatrix} 11/28 & 17/84 & 17/84 & 17/84 \\ 17/84 & 11/28 & 17/84 & 17/84 \\ 17/84 & 17/84 & 11/28 & 17/84 \\ 17/84 & 17/84 & 17/84 & 11/28 \end{bmatrix} \quad (11)$$

This matrix satisfies the conditions of efficiency balanced design, i.e. $\mathbf{MJ} = \mathbf{J}$, where $\mathbf{J}$ is the $(v \times 1)$ unit vector.

The efficiency factor is calculated by using the formula

$$\mathbf{M} = \mu \mathbf{I} + (1 - \mu) \mathbf{J} \mathbf{r}' / n$$

$$\mathbf{M} = \begin{bmatrix} \mu & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & \mu \end{bmatrix} + \left\{ (1 - \mu) \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} [7 \ 7 \ 7 \ 7] \right\} / 28$$

$$\mathbf{M} = \begin{bmatrix} \mu & 0 & 0 & 0 \\ 0 & \mu & 0 & 0 \\ 0 & 0 & \mu & 0 \\ 0 & 0 & 0 & \mu \end{bmatrix} + \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} - \mu \begin{bmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{bmatrix} \quad (12)$$

Equating (11) and (12) we get $\mu = \frac{4}{21}$.

The design with the above incidence matrix gives an efficiency balanced design with parameters

$$v = 4, b = 11, \mathbf{r} = (7 \ 7 \ 7 \ 7) \text{ and } \mathbf{k} = \left(\underbrace{2 \ 2 \ 2 \ 2 \ 2}_{4C_2 \text{ times}} \underbrace{23 \ 3 \ 3 \ 3}_{4C_3 \text{ times}} \underbrace{4}_{4C_4 \text{ times}} \right)$$

having efficiency factor $E = (1 - \mu) = \frac{17}{21}$.

References

- Awad, R., & Banerjee, S. (2012). Some construction methods of variance and efficiency balanced block designs with repeated blocks. *Research Journal of Recent Sciences*, 1(9), 18-22. Retrieved from <http://www.isca.in/rjrs/archive/v1i9/4.ISCA-RJRS-2012-176.pdf>
- Calinski, T. (1971). On some desirable patterns in block designs. *Biometrics*, 27, 275-292. doi: 10.2307/2528995
- Ceranka, B., & Graczyk, M. (2009). Some notes about efficiency balanced block designs with repeated blocks. *Metodološki Zvezki*, 6(1), 69-76. Retrieved from <http://www.stat-d.si/mz/mz6.1/ceranka1.pdf>

CONSTRUCTION OF EFFICIENCY BALANCED DESIGN

Dey, A., Singh, M., & Saha, G.M. (1981). Efficiency balanced block designs. *Communications in Statistics - Theory and Methods*, 10(3), 237-247. doi: 10.1080/03610928108828034

Gupta, S. (1992). Efficiency balance through BIB and GD designs. *The Indian Journal of Statistics, B*, 54(2), 220-226. Available from <http://www.jstor.org/stable/25052739>

Gupta, V. K. & Gandhi, P. N. S. (1991). On construction of general efficiency balanced block designs. *The Indian Journal of Statistics, B*, 53(1), 89-96. Available from <http://www.jstor.org/stable/25052680>

Hedayat, A., & Federer, W. T. (1974). Pairwise and variance balanced incomplete block design. *Annals of the Institute of Statistical Mathematics*, 26(1), 331-338. doi:10.1007/BF02479828

Jones, R. M. (1959). On a property of incomplete blocks. *Journal of the Royal Statistical Society: Series B*, 21, 172-79. Available from <http://www.jstor.org/stable/2983939>

Mukerjee, R., & Saha, G. M. (1990). Some optimality results on efficiency-balanced designs. *The Indian Journal of Statistics, B*, 52(3), 324-331. Available from <http://www.jstor.org/stable/25052659>

Puri, P. D. & Nigam, A. K. (1975). A note on efficiency balanced designs. *The Indian Journal of Statistics, B*, 37(4), 457-460. Available from <http://www.jstor.org/stable/25051981>

Sun, T. & Tang, Y. (2010). Optimal efficiency balanced designs and their constructions. *Journal of Statistical Planning and Inference*, 140, 2771-2777. doi: 10.1016/j.jspi.2010.03.039

Williams, E. R. (1975). Efficiency-balanced designs. *Biometrika*, 62, 686-689. Available from <http://www.jstor.org/stable/2335531>

Semiparametric Mixed Models for Nested Repeated Measures Applied to Ambulatory Blood Pressure Monitoring Data

Rhonda D. Szczesniak
Cincinnati Children's Hospital
Cincinnati, OH

Dan Li
University of Cincinnati
Cincinnati, OH

Raouf S. Amin
Cincinnati Children's Hospital
Cincinnati, OH

Semiparametric mixed models are increasingly popular for statistical analysis of medical device studies in which long sequences of repeated measurements are recorded. Monitoring these sequences at different periods over time on the same individual, such as before and after an intervention, results in nested repeated measures (NRM). Covariance models to account for NRM and simultaneously address mean profile estimation with penalized splines via semiparametric regression are considered with application to a prospective study of 24-hour ambulatory blood pressure and the impact of surgical intervention on obstructive sleep apnea.

Keywords: functional data analysis, semiparametric regression, blood pressure, longitudinal data analysis, doubly repeated measures, obstructive sleep apnea, medical monitoring, circadian rhythm

---

Introduction

Medical device studies frequently involve collections of multiple recordings that result in long sequences of repeated measurements for each subject. It is often of interest to assess these sequences at different periods of time or recording sessions on the same subject. This type of data, commonly called nested repeated measures (NRM), yields two sources of intrasubject variation: an inner source arising from observations within a sequence and an outer source arising from observations under different time periods, such as before and after an intervention. Covariance models for NRM have been proposed to account for the intrasubject correlation arising from data of this nature (Harville, 1997; Laird & Ware, 1982; Jennrich & Schluchter, 1986). More recent work (Park & Lee, 2002) shows covariate effects are impacted by the choice of covariance structure and a series of covariance

---

Dr. Szczesniak is an Associate Professor in the Division of Biostatistics & Epidemiology. Email her at: rhonda.szczesniak@cchmc.org.

models to account for NRM can be specified in a linear mixed model. In these applications, mean profiles are modeled using polynomial regressions. The models are used to compare different experimental conditions or mean profiles by assessing overall mean differences.

For experiments in which the mean profile cannot be characterized with a parametric function, semiparametric mixed models may be useful. Penalized splines are a more flexible alternative to estimate the mean function (Eilers & Marx, 1996), and can be expressed as the best linear unbiased prediction (BLUP) of a linear mixed model (Ruppert, Wand, & Carroll, 2003, p. 99-100). Combining this nonparametric representation of the mean function with parametric estimation in the linear mixed model is referred to as semiparametric regression. Semiparametric mixed models have been used to compare the mean profiles of two independent groups (e.g. placebo versus treated) in a study of cardiovascular safety data (Maringwa et al., 2008a). Model selection was performed and mean profiles were estimated with linear penalized splines. The group-specific mean profiles were compared over time using simultaneous confidence bands. This approach, which has been used in other biomedical studies (see VanDyke et al., 2012 for an example), works well for single repeated measures factors. Semiparametric mixed models have also been applied to data arising from crossover designs to compare condition-specific mean profiles over time (Maringwa et al., 2008b). In this study focused on crossover designs, correlation between and within periods were assumed to be separable (Jones & Kenward, 2003, p. 193). This assumption corresponds to concluding that the outer repeated measures (between periods) may be accounted for by using subject-specific random intercepts. In NRM studies where outer repeated measures are collected at variable times across subjects, a more complex correlation structure may be necessary.

Despite inferential goals for time-specific comparisons that are similar to prior developments, the combination of NRM and an unrecognizable mean response function requires further methodological development for efficient regression parameter estimates. In this article, a series of semiparametric mixed models are proposed which incorporate NRM covariance modeling and mean profile estimation approaches. The following section begins with description of the ambulatory blood pressure monitoring data, which is the motivation for this development. Model selection criteria are then provided along with calculations for simultaneous confidence bands to assess time-specific intervention effects with application to the motivating data. The appropriateness of each proposed model for the data is discussed. Additional details on covariance models and relevant code are provided as Appendices.

Methodology

Twenty-Four Hour Ambulatory Blood Pressure Monitoring

The motivating data arises from a prospective study to examine the effect of a surgical intervention on 24-hour ambulatory blood pressure. The importance of diurnal changes in blood pressure in predicting target organ damage has been demonstrated (Mansoor & Massie, 1999). The application in this article focuses on diastolic blood pressure (DBP) profiles arising from ambulatory blood pressure monitoring. These profiles are recordings of systolic and diastolic blood pressures, observed in 30-minute intervals over a 24-hour period beginning with time of sleep onset for each subject. In healthy subjects, these data typically have a marked circadian pattern with diurnal features that may not be present in subjects with obstructive sleep apnea (Mansoor, 2002). Previous analysis techniques have included the use of restricted cubic splines to fit DBP profiles in a study of hypertension during pregnancy (Lambert, Abrams, Jones, Halligan, & Shennan, 2001).

This application focuses on whether DBP patterns change in subjects with obstructive sleep apnea after an intervention consisting of adenotonsillectomy. For the study, each subject wore an ambulatory blood pressure monitoring cuff at two separate recording sessions: before and after the intervention. The duration between baseline and follow-up ranged from six to twelve months and warranted consideration of more complex covariance models, as any level of improvement after intervention may be time-sensitive. Immediately prior to receiving the cuff to monitor blood pressure, demographic and clinical characteristics were obtained for each subject. Although monitoring data were equally spaced and collected both before and after the intervention for each of the 58 subjects, there were incomplete profiles due to cuff malfunctions that occurred during each 24-hour observation period. The timing of the follow-up monitoring and potential for incomplete profiles increase the importance of selecting an appropriate covariance model.

Figure 1 illustrates the pre- and post-surgical intervention profiles of five randomly selected subjects, demonstrating the intrasubject variability arising from the inner-repeated measures (within profiles) and the outer repeated measures (baseline and follow-up profiles), as well as the intersubject variability between profiles. Circadian rhythm in daytime and nighttime blood pressures is not apparent in the observed profiles, presumably due to the intra- and intersubject variation. In addition to assessing intervention effects on mean DBP response, the rate of change

Figure 1. Subject-specific observed DBP response, before and after intervention

Five subject-specific profiles during recording of 24-hour diastolic blood pressure (DBP, y-axis) taken prior to surgery (left) and after surgery (right); x-axis represents time (in hours) since sleep onset

experienced by subjects when awakening from sleep (clinically termed the “morning blood pressure surge”) is also of clinical importance. The presence of this feature indicates greater subject responsiveness to blood pressure regulation. Rates of change close to zero may indicate poor blood pressure control ([Amin et al., 2008](#); [Crisalli et al., 2012](#)).

The statistical methodology in this manuscript relates to three inferential goals. First, develop an appropriate model for NRM covariance and spline representation of baseline and follow-up AMBP profiles. Second, determine how the rate of change or “morning surge” changes over time by using first-order derivatives of penalized regression splines. Third, construct simultaneous confidence bands to compare mean differences between baseline and follow-up AMBP profiles during daytime and nighttime.

Modeling the Mean Response Function

A penalized spline representation ([Eilers & Marx, 1996](#)) is used to model the mean DBP response over the 24-hour interval. Ignoring intervention effect, this model can be expressed as

$$\begin{aligned} \text{DBP}_{ijk} &= f(t_{ijk}) + \varepsilon_{ijk}^*, \\ f(t_{ijk}) &= \beta_0 + \beta_1 t_{ijk} + \dots + \beta_p t_{ijk}^p + \sum_{l=1}^L b_l (t_{ijk} - \kappa_l)_+^p \end{aligned} \quad (1)$$

The term DBP_{ijk} represents the measurement of DBP for the i^{th} subject ($i = 1, \dots, n$) taken at the j^{th} recording session ($j = 1, \dots, n_i$) at time t_{ijk} ($k = 1, \dots, n_{ij}$) during the 24-hour interval. The function $f(t_{ijk})$ is a combination of fixed effects parameters $(\beta_0, \beta_1, \dots, \beta_p)'$ and p^{th} degree splines with knots at distinct locations $(\kappa_1, \dots, \kappa_L)'$ along the time interval with corresponding coefficients $(b_1, \dots, b_L)'$; assume that $b_l \sim N(0, \sigma_b^2)$. Representations for the error term ε_{ijk}^* are discussed in the context of the linear mixed model later in this section. A series of models similar to those from previous work using linear truncated power splines (Maringwa et al., 2008a) but expanded to incorporate NRM arise from (1) and may be considered to represent the overall shape of the 24-hour DBP, (Table 1). The knot locations are in the range of t_{ijk} values, where $t_+ = \max(0, t)$. To fit the mean function in our motivating example, quadratic ($p = 2$) penalized splines were used.

Structure (1.1) in Table 1 shows a common DBP curve for both pre- and post-intervention, corresponding to no intervention effect. To assess whether the intervention effect is parallel, one can examine Structure (1.2). It is possible that post-intervention profiles have an average quadratic trend that differs from the trend during pre-intervention, without any changes to the more localized, nonparametric (spline) portion of the model. For this case, one can examine Structure (1.3). One can fit Structure (1.4) to capture more localized changes in average DBP features. The above distribution and independence assumptions hold in all model settings, except Structure (1.5.), which provides different degrees of smoothing based on whether the session occurred before or after intervention.

Semiparametric regression and NRM can be characterized in the familiar linear mixed model framework for longitudinal data (Verbeke & Molenberghs, 2000, p. 23) as

$$\mathbf{Y}_i = \mathbf{X}_i \boldsymbol{\beta}_i + \mathbf{Z}_i \mathbf{b}_i + \boldsymbol{\varepsilon}_i^*, \quad (2)$$

where $\mathbf{X}_i$ and $\boldsymbol{\beta}_i$ represent the traditional fixed effects design matrix and parameter vector, $\mathbf{Z}_i$ and $\mathbf{b}_i$ correspond to the previously-described design matrix for the spline basis function and coefficient vector, and the overall error vector $\boldsymbol{\varepsilon}_i^*$ corresponds to the ε_{ijk}^* in (1); $\boldsymbol{\varepsilon}_i^* \sim N(0, \boldsymbol{\Sigma}_i^*)$. The entire response for the i^{th} subject,

$\mathbf{Y}_i$, is an $n_i \times 1$ vector, where $n_i = \sum_{j=1}^{n_i} n_{ij}$.

NESTED SEMIPARAMETRIC MIXED EFFECTS MODELS

Table 1. Mean response models for twenty-four hour diastolic blood pressure

Effects Description	Mean Response Structure*
(1.1) No intervention effect	$\beta_0 + \beta_1 t_{ijk} + \beta_2 t_{ijk}^2 + \sum_{l=1}^L b_l (t_{ijk} - \kappa_l)_+^2$
(1.2) Intervention effect constant across 24-hour sequence	$\beta_0 + \beta_{02} \text{post}_{ij} + \beta_1 t_{ijk} + \beta_2 t_{ijk}^2 + \sum_{l=1}^L b_l (t_{ijk} - \kappa_l)_+^2$
(1.3) Pre- and post- intervention profiles have different quadratic trends	$\beta_0 + \beta_{02} \text{post}_{ij} + \beta_1 t_{ijk} + \beta_{12} \text{post}_{ij} t_{ijk} + \beta_2 t_{ijk}^2 + \beta_{12} \text{post}_{ij} t_{ijk}^2 + \sum_{l=1}^L b_l (t_{ijk} - \kappa_l)_+^2$
(1.4) Pre- and post-intervention profiles smoothed differently using distinct vectors for coefficients for pre- and post-intervention profiles	$\beta_0 + \beta_{02} \text{post}_{ij} + \beta_1 t_{ijk} + \beta_{12} \text{post}_{ij} t_{ijk} + \beta_2 t_{ijk}^2 + \beta_{12} \text{post}_{ij} t_{ijk}^2 + \sum_{l=1}^L b_l^{\text{pre}_{ij}} (t_{ijk} - \kappa_l)_+^2 + \sum_{l=1}^L b_l^{\text{post}_{ij}} (t_{ijk} - \kappa_l)_+^2$
(1.5) Separate smoothing and distinct smoothing parameters for pre- and post-intervention profiles	Same as structure (1.4) but differing variances for smoothing coefficients: $b_l^{\text{pre}_{ij}} \sim N(0, \sigma_{b_l^{\text{pre}}}^2)$ and $b_l^{\text{post}_{ij}} \sim N(0, \sigma_{b_l^{\text{post}}}^2)$

* The term post_{ij} refers to an indicator of post-intervention assessment (1 if observation taken during post-intervention session, 0 otherwise); pre_{ij} is defined similarly for pre-intervention

The proposed structure provides flexibility for the covariance matrices, which may be advantageous for NRM. As described by Park and Lee (2002), the subject-specific covariance matrix for the error term $\boldsymbol{\varepsilon}_i^*$ is

$$\boldsymbol{\Sigma}_i^* = \begin{bmatrix} \text{Var}(\boldsymbol{\varepsilon}_{i1}^*) & \text{Cov}(\boldsymbol{\varepsilon}_{i1}^*, \boldsymbol{\varepsilon}_{i2}^*) & \cdots & \text{Cov}(\boldsymbol{\varepsilon}_{i1}^*, \boldsymbol{\varepsilon}_{in_i}^*) \\ \text{Cov}(\boldsymbol{\varepsilon}_{i1}^*, \boldsymbol{\varepsilon}_{i2}^*) & \text{Var}(\boldsymbol{\varepsilon}_{i2}^*) & \cdots & \text{Cov}(\boldsymbol{\varepsilon}_{i2}^*, \boldsymbol{\varepsilon}_{in_i}^*) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(\boldsymbol{\varepsilon}_{i1}^*, \boldsymbol{\varepsilon}_{in_i}^*) & \text{Cov}(\boldsymbol{\varepsilon}_{i1}^*, \boldsymbol{\varepsilon}_{i2}^*) & \cdots & \text{Var}(\boldsymbol{\varepsilon}_{in_i}^*) \end{bmatrix} \quad (3)$$

The $\text{Var}(\boldsymbol{\varepsilon}_{ij}^*)$ is the $n_i \times n_i$ variance-covariance matrix for the sequence of measurements from the i^{th} subject observed on the j^{th} occasion; $\text{Cov}(\boldsymbol{\varepsilon}_{ij}^*, \boldsymbol{\varepsilon}_{ij'}^*)$ for $j \neq j'$ is the covariance matrix for measurements observed at distinct occasions j and j' . Different covariance models are now presented for the motivating example.

Covariance for Nested Repeated Measures

Random effects: Random intercepts are one of the most commonly used methods to address intrasubject variability. In the case of a single repeated factor, subject-specific effects are often included as random intercepts and all measurements are assumed to have equal correlation. This assumption corresponds to compound symmetry. The approach can be naturally extended to NRM by including two additional random effects for occasion and sequence. For a given subject, any two measurements taken within a sequence during a single occasion have correlation ρ_s ; two measurements taken at the same time point of sequences on two distinct occasions have correlation ρ_o . This covariance model has straightforward interpretation but may not be suited for many experiments with NRM. In the motivating DBP example, ρ_s corresponds to an individual's measurements taken within a 24-hour period having the same correlation, regardless of the amount of time lapsing between measurements; the outer repeated measure correlation ρ_o assumes that any two DBPs recorded at the same time during two separate 24-hour periods have a common correlation.

Composite covariance: Nonconstant correlation within the 24-hour period and unequal variances between visits are both plausible but neither can be addressed with the aforementioned random effects covariance structure. Instead, one can use a composite covariance model (Searle, 2006, p. 348) obtained using the right Kronecker product to model the sources of correlation arising from inner and outer repeated measures. There are several possibilities for composite covariance models, although there are some limitations imposed by software capabilities (Park & Lee, 2002). For the DBP example, an unstructured covariance for the outer repeated measure (occasion) and AR(1) structure for the inner repeated measure (sequence) are considered. The covariance matrix corresponding to the direct product of unstructured and AR(1) covariance is

$$\begin{aligned}
 & \begin{bmatrix} \sigma_1^2 & \sigma_{21} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix} \otimes \begin{bmatrix} 1 & \rho & \rho^2 & \dots \\ \rho & 1 & \rho & \rho^2 & \dots \\ \rho^2 & \rho & 1 & & \\ \vdots & \vdots & & \ddots & \\ & & & & 1 \end{bmatrix} \\
 &= \begin{bmatrix} \sigma_1^2 & \sigma_1^2 \rho & \sigma_1^2 \rho^2 & \dots & \sigma_{21} & \sigma_{21} \rho & \sigma_{21} \rho^2 & \dots \\ \sigma_1^2 \rho & \sigma_1^2 & \sigma_1^2 \rho & \sigma_1^2 \rho^2 & \sigma_{21} \rho & \sigma_{21} & \sigma_{21} \rho & \sigma_{21} \rho^2 \\ \sigma_1^2 \rho^2 & \sigma_1^2 \rho & \sigma_1^2 & & \sigma_{21} \rho^2 & \sigma_{21} \rho & \sigma_{21} & \\ \vdots & \vdots & & \ddots & \vdots & \vdots & & \ddots \\ & & & & \sigma_1^2 & & & \sigma_{21} \\ \sigma_{21} & \sigma_{21} \rho & \sigma_{21} \rho^2 & \dots & \sigma_2^2 & \sigma_2^2 \rho & \sigma_2^2 \rho^2 & \dots \\ \sigma_{21} \rho & \sigma_{21} & \sigma_{21} \rho & \sigma_{21} \rho^2 & \sigma_2^2 \rho & \sigma_2^2 & \sigma_2^2 \rho & \sigma_2^2 \rho^2 \\ \sigma_{21} \rho^2 & \sigma_{21} \rho & \sigma_{21} & & \sigma_2^2 \rho^2 & \sigma_2^2 \rho & \sigma_2^2 & \\ \vdots & \vdots & & \ddots & & & & \ddots \\ & & & & \sigma_{21} & & & \sigma_2^2 \end{bmatrix}
 \end{aligned}$$

Exponential covariance: The composite covariance model assumes equally spaced measurement times within each occasion (at the sequence level) and between occasions. In the motivating example, the timing of post-intervention measurement is not equal across subjects, which suggests the need for a more flexible covariance model. The exponential covariance models described in this section have their origin in spatial statistics but can be used in the linear mixed model framework (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006, p. 198) and have been used in various longitudinal data analysis applications with a single repeated measures factor (see Szczesniak et al., 2013 for a recent biomedical example). This section covers two general types of exponential covariance models.

Consider the semivariogram formula for the exponential covariance model with nugget effect:

$$r(h) = C_n + \sigma_0^2 \left[1 - \exp \left(- \frac{|t_{ijk} - t_{ijk'}|}{a_0} \right) \right],$$

where $|t_{ijk} - t_{ijk'}|$ refers to the Euclidean distance between two time points, t_{ijk} and $t_{ijk'}$, thereby relaxing the previously-described assumption of equally-spaced inner repeated measures for the AR(1) covariance and accounting for irregular time series. The terms C_n , $C_n + \sigma_0^2$, and a_0 correspond to the geostatistical parameters referred to as the nugget, sill, and range (Wackernagel, 2003, p. 57). The nugget effect is a measure of the residual error or white noise of the DBP response; the range parameter dictates the decay of the covariance function. Please see Appendix A for a graphical explanation of these terms. If the term C_n is excluded, then the model is considered an exponential covariance model without the nugget effect. Both versions of this exponential covariance model are applied to the motivating example.

Results

Fifty-eight subjects completed both recording sessions, and their observations comprise the data of interest. The median (Q1-Q3) time between pre- and post-intervention measurement periods was 288.3 (218.1-321.5) days and ranged from as few as 177 days between visits to as much as 364 days between visits. This indicates the potential need to model unequally-spaced repeated measurements. The number of observed half-hourly DBP recordings over the 24-hour period was 41.0 (34.5-44.8) and 42.4 (36.0-46.2) at the baseline and follow-up sessions, respectively. Baseline age and BMI z-score were 9.0 (7.1-11.5) years and 1.47 (0.34-2.25), respectively; 40.4% of subjects were Caucasian and 46.8% were male.

The series of models from Table 1 were used to characterize DBP over the 24-hour sequence and the two measurement occasions. The mean response model chosen using adjusted fit statistics presented in previous work (Maringwa et al., 2008a) had Structure (1.3), which provided separate polynomial terms for each occasion but relied on the same smoothing parameter.

Each of the four previously described covariance functions was applied to model the correlation for the ϵ_i^* term in (2). SAS code to implement the covariance models are in Appendix B. The estimates for the demographic covariates under each covariance model are presented in Table 2. Effect estimates were consistent across the four different covariance models, except for the effect of race, but this effect was not statistically significant in any of the models. Gender and BMI z-score were statistically significant in all models.

NESTED SEMIPARAMETRIC MIXED EFFECTS MODELS

Table 2. Linear covariate effects* estimates from covariance models of diastolic blood pressure

Effect	Covariance Model			
	Random Effects Estimate (SE) p-value	Composite Covariance Estimate (SE) p-value	Exponential (with nugget effect) Estimate (SE) p-value	Exponential (without nugget effect) Estimate (SE) p-value
Male	3.13970 (1.10020) 0.00640	3.14310 (1.11280) 0.00690	3.13940 (1.12850) 0.00780	3.19640 (1.10690) 0.00590
Caucasian	-0.08742 (1.09580) 0.93680	-0.08212 (1.10810) 0.94120	-0.03037 (1.12400) 0.97860	0.01099 (1.10250) 0.99210
Age (years)	0.18210 (0.22670) 0.42570	0.19510 (0.23100) 0.40210	0.18380 (0.23670) 0.44100	0.17270 (0.23080) 0.45780
Body Mass Index (z-score)	1.47780 (0.48580) 0.00290	1.49850 (0.51540) 0.00460	1.51340 (0.56210) 0.00900	1.55500 (0.52990) 0.00440

* These effects were assumed to enter the models linearly. Each model included the mean response function specified in Table 1, Structure (1.3)

Evaluating Model Fit

An important task in the model-building process is to select a suitable covariance structure. The effective number of parameters, referred to as E_p , can be obtained for each covariance model by estimating the appropriate covariance structure using (3). Let $\mathbf{C} = [\mathbf{X} \mathbf{Z}_b]$ be the design matrix for the mean response function and $\mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G}^{-1} \end{bmatrix}$, where $\mathbf{G}$ corresponds to any random effects being used to model

covariance (e.g. random intercepts), $\mathbf{R} = \text{blkdiag}(\boldsymbol{\epsilon}_i^*)$, $i = 1, \dots, n_i$. The E_p for each covariance model may be computed as: $E_p = \text{trace}((\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1} \mathbf{C}^T \mathbf{R}^{-1} \mathbf{C})$.

Ultimately, the adjusted AIC can be computed as $\text{AIC}_{\text{adj}} = -2LL + 2E_p$. This calculation will take into account the additional parameters brought about by fitting the mean response function $f(t)$ and the covariance function. SAS code for the E_p calculations necessary for model structures in Table 1 and NRM covariance models is available from the authors upon request.

The fit statistics were calculated for each covariance model and are displayed in Table 3. The adjusted AIC and the more common information criteria (both marginal AIC and BIC) indicated that exponential covariance with a nugget effect provides the best fit of the covariance models considered. Subsequent estimation for the intervention effect is based on the exponential covariance with nugget.

Table 3. Fit statistic results for covariance models

Covariance Model*	Fit Statistics				
	-2loglikelihood	AIC	BIC	Ep	AlCadJ
Random Effects	33494.0	33520.0	33494.0	12.1291	33518.3
Composite Covariance	33264.9	33296.9	33296.9	12.0024	33288.9
Exponential (with nugget)	32874.1	32904.1	32874.1	12.0604	32898.2
Exponential (without nugget)	33042.5	33070.5	33042.5	11.6404	33065.8

* Each covariance model includes the mean response function specified in Table 1, Structure (1.3)

Figure 2. Fitted DBP response curves and morning surge, before and after intervention

The averaged response (jagged line), fitted curve (smooth, solid line) for $f(t)$ and corresponding 95% simultaneous confidence bands (dashed lines) during recording of the 24-hour diastolic blood pressure (DBP, y -axis) taken prior to surgery in (a) and after surgery in (b). The plot in (c) shows the derivative of the smooth function $f(t)$ for pre- and post-intervention sessions of 24-hour DBP recordings, where the solid (dashed) curve represents the rate of change for the pre-intervention (post-intervention) recordings. The difference between 24-hour DBP mean response functions before and after intervention (solid line) and 95% simultaneous confidence bands (dashed lines) are presented in (d).

Figure 2 (a-b) shows the fitted DBP curves corresponding to occasions before and after intervention, which both show smoother trends than the averaged response. The 95% confidence bands accompanying each fitted curve are explained in a subsequent section. The circadian rhythm is more discernable here than examining the individual functions in Figure 1, and corresponds to previously mentioned studies that suggest a diurnal response over time. Nocturnal dipping is a feature that indicates healthier DBP rhythm. It is present at both occasions and is noticeable just before the sleep cycle starts for the next day (around $t = 22$ hours after sleep onset).

The derivative $f'(t)$ of Structure (1.3) in Table 1 can be used to examine the morning surge before and after intervention. By looking at time since sleep onset, which corresponds to 0 on the x -axis in Figure 2 (c), the rate of change in average DBP is slightly higher for the intervention period; however, the derivative curves begin to overlap around $t = 12$ hours after sleep onset. Presumably, 7-9 hours after sleep onset is the interval of interest to assess the morning surge, as this is the time frame when subjects begin to wake. Results suggest the rate of change is slightly elevated after intervention, as compared to before intervention. From a biomedical perspective, this finding may indicate heightened response to wakefulness as a result of receiving the intervention.

Simultaneous Confidence Bands

It is also of interest to examine the intervention effect on mean DBP response across the 24-hour interval. It is plausible to conduct point-by-point comparisons of the occasion-specific mean response functions. Rather than making this comparison of $f_{\text{pre}}(t)$ to $f_{\text{post}}(t)$ for all observed t in the 24-hour interval, one can avoid those multiple comparison issues by constructing a simultaneous confidence band for this difference by using the following result (Ruppert et al., 2003, p. 142-143):

$$\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_b - \mathbf{b}_b \end{bmatrix} \sim N \left\{ \mathbf{0}, \left(\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B} \right)^{-1} \right\} \quad (4)$$

Define a grid $\mathbf{g}$ of time points $(0, 23)$ by increments of 0.5 hours such that there are $T = 49$ equally spaced time points $(g_1, \dots, g_{49})'$. One can evaluate the estimated difference between the two functions as

$$\hat{\mathbf{f}}_d - \mathbf{f}_d = (\hat{\mathbf{f}}_{\text{pre}} - \hat{\mathbf{f}}_{\text{post}}) - (\mathbf{f}_{\text{pre}} - \mathbf{f}_{\text{post}}) = \mathbf{C}_g \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_b - \mathbf{b}_b \end{bmatrix} = [\mathbf{L}\mathbf{X} \quad \mathbf{L}\mathbf{Z}_b] \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_b - \mathbf{b}_b \end{bmatrix},$$

where $\mathbf{L} = \begin{bmatrix} 1 & & -1 & \\ & \ddots & & \ddots \\ & & 1 & -1 \end{bmatrix}_{49 \times 98}$ is a contrast matrix, $\mathbf{X}$ and $\mathbf{Z}_b$ are the design matrices evaluated over $\mathbf{g}$, and $\mathbf{C}_g = [\mathbf{L}\mathbf{X} \quad \mathbf{L}\mathbf{Z}_b]$.

In order to obtain the $\text{stdev}[\hat{\mathbf{f}}_d - \mathbf{f}_d]$ for the confidence band, it is necessary to compute $\text{Cov}[\hat{\mathbf{f}}_d - \mathbf{f}_d]$. Using the following result:

$$\text{Cov}\left(\begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_b - \mathbf{b}_b \end{bmatrix}\right) = (\mathbf{C}^T \mathbf{R}^{-1} \mathbf{C} + \mathbf{B})^{-1},$$

one can obtain a 95% simultaneous confidence band for $\mathbf{f}_d$ as

$$\left[\hat{\mathbf{f}}_d \pm h_{0.95} \text{stdev}\{\hat{\mathbf{f}}_d - \mathbf{f}_d\} \right]_{1 \leq l \leq T},$$

where $h_{0.95}$ is the $1 - \alpha$ quantile with $\alpha = 0.05$. Finally, it can be approximated as

$$\sup \left| \frac{\hat{\mathbf{f}}_d - \mathbf{f}_d}{\text{stdev}\{\hat{\mathbf{f}}_d - \mathbf{f}_d\}} \right| \approx \max_{1 \leq l \leq T} \left| \frac{\left(\mathbf{C}_g \begin{bmatrix} \hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \\ \hat{\mathbf{b}}_b - \mathbf{b}_b \end{bmatrix} \right)_l}{\text{stdev}\{\hat{\mathbf{f}}_d - \mathbf{f}_d\}} \right|. \quad (5)$$

As an example, if simulations from (4) then computations of (5) are repeated 10,000 times, then the value of the ranked 9,500th quantity is used as $h_{0.95}$. Similarly, a 95% pointwise confidence band for $\mathbf{f}_d$ is

$$\left[\hat{\mathbf{f}}_d \pm Z_{0.95} \text{stdev}\{\hat{\mathbf{f}}_d - \mathbf{f}_d\} \right]_{1 \leq l \leq T},$$

where $Z_{0.95} = 1.96$. SAS code for both simultaneous and pointwise confidence bands is available from the authors upon request.

The simultaneous confidence band for $\mathbf{f}_d$ in the DBP example is plotted in Figure 2 (d). Portions of the band that do not overlap with zero on the y -axis are considered to indicate statistically significant differences between pre- and post-intervention periods. As expected, the significant differences occur roughly for the first 5 hours of sleep (from $t = 0$ to $t = 5$ hours) and indicate mean DBP lowers after intervention. The mean profiles otherwise show substantial overlap, particularly during daytime ($t > 10$ hours). If there is any effect from intervention, it likely occurs during nighttime. Although not shown here, pointwise confidence bands for $\mathbf{f}_d$ indicate similar findings but have narrower bands because there is no adjustment for simultaneity.

Conclusion

With so many devices offering the opportunity to measure real-time subject outcomes over extended periods of time, many researchers may be overwhelmed by the amount of data and the task of determining an appropriate statistical method to assess treatment effects. Extending semiparametric mixed models to account for NRM offers a solution to such challenges. In the motivating example with 24-hour DBP recording, using this approach shows that intervention effects may be observable during sleep. It is likely that these findings would be masked if one tests summary measures from the DBP curves. Incorporating penalized splines provided a more sensitive means to assess medically important features of the DBP profile, such as nocturnal dipping and morning surge. Findings using semiparametric regression suggest the presence of an unexpected “daytime dip.” These findings are not consistent with the DBP profiles of healthy controls but reflect prior studies of rough averages of DBP over the 24-hour period (Amin et al., 2008).

Ignoring the impact from NRM on the regression model reduces efficiency in the parameter estimates and may lead to incorrect conclusions about intervention effects. Analyses of the DBP data show improvement in model fit is attributable to accounting for unequally spaced measurement times. There are other covariance models that can also be implemented in the SAS MIXED procedure to account for NRM. Some examples include the Gaussian covariance model (nugget effect specification is optional) and the right Kronecker product “AR(1)⊗UN,” which corresponds to the DBP data to having AR(1) covariance for the outer repeated measure (occasion) and unstructured covariance for the inner repeated measure (sequence). The authors attempted to fit these covariance models to the data but

estimates were not obtained due to convergence issues from the MIXED procedure. An alternative approach to the semiparametric mixed models presented here is to perform sequential or hierarchical regression via path analysis (Snijders, 1996). Additional consideration in the model setup would be needed to incorporate the spline basis functions at subject-specific levels.

There are several ways in which the semiparametric mixed model with NRM covariance presented here can be further explored and extended. Functional principal components analysis may be used to examine dominant modes of variation in the subject- and visit-specific DBP profiles (Silverman, 1996); recent developments have been made to apply this approach on NRM (Shou, Zipunniokov, Crainiceanu, & Greven, 2014). If data have a mean response function with sharp changes, multiple knots may be desirable in that region, and it may be advantageous to change knot locations of sequences observed at different periods. For such instances, adaptive spline methods may be useful; however, some methods may require different estimation approaches (DiMatteo, Genovese, & Kass, 2001). It may also be of interest to assess the correlation between spline coefficients for the difference between occasions. Clinical and demographic characteristics in this study were assumed to enter the model linearly as covariates, but that assumption may be relaxed using generalized additive models (Hastie & Tibshirani, 1990, p. 136-171). Although not the focus of this work, missing data is a pervasive issue. The approach used in this study essentially assumed the missing mechanism was MAR (Rubin, 1976); however, more recent work has been done to improve efficiency of estimators in semiparametric regression models in the presence of missing data (Yu & Nan, 2006). That work may be extended to the NRM covariance models presented here.

References

- Amin, R., Somers, V. K., McConnell, K., Willging, P., Myer, C., Sherman, M., McPhail, G. L., Morgenthal, M., Fenchel, M. C., Bean, J., Kimball, T., & Daniels, S. (2008). Activity-adjusted 24-hour ambulatory blood pressure and cardiac remodeling in children with sleep disordered breathing. *Hypertension*, *51*(1), 84-91. doi: [10.1161/HYPERTENSIONAHA.107.099762](https://doi.org/10.1161/HYPERTENSIONAHA.107.099762)
- Crisalli, J. A., McConnell, K., VanDyke, R. D., Fenchel, M. C., Somers, V. K., Shamsuzzaman, A., Chini, B., Daniels, S. R., & Amin, R. (2012). Baroreflex sensitivity after adenotonsillectomy in children with obstructive sleep apnea during wakefulness and sleep. *Sleep*, *35*(10), 1335-1343. doi: [10.5665/sleep.2108](https://doi.org/10.5665/sleep.2108)

DiMatteo, I., Genovese, C. R., & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4), 1055-1071. doi:

[10.1093/biomet/88.4.1055](https://doi.org/10.1093/biomet/88.4.1055)

Eilers, P. H. C., & Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2), 89-102. doi: [10.1214/ss/1038425655](https://doi.org/10.1214/ss/1038425655)

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358), 320-338. doi: [10.1080/01621459.1977.10480998](https://doi.org/10.1080/01621459.1977.10480998)

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Monographs on statistics and applied probability. Boca Raton, FL: Chapman & Hall/CRC.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42(4), 805-820. doi: [10.2307/2530695](https://doi.org/10.2307/2530695)

Jones, B., & Kenward, M. G. (2003). *Design and analysis of cross-over trials* (2nd ed.). Monographs on statistics and applied probability. Boca Raton, FL: Chapman & Hall/CRC.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4), 963-974. doi: [10.2307/2529876](https://doi.org/10.2307/2529876)

Lambert, P. C., Abrams, K. R., Jones, D. R., Halligan, A. W., & Shennan, A. (2001). Analysis of ambulatory blood pressure monitor data using a hierarchical model incorporating restricted cubic splines and heterogeneous within-subject variances. *Statistics in Medicine*, 20(24), 3789-3805. doi: [10.1002/sim.1172](https://doi.org/10.1002/sim.1172)

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models*. Cary, NC: SAS Institute, Inc.

Mansoor, G. A. (2002). Sleep actigraphy in hypertensive patients with the 'non-dipper' blood pressure profile. *Journal of Human Hypertension*, 16(4), 237-242. doi: [10.1038/sj.jhh.1001383](https://doi.org/10.1038/sj.jhh.1001383)

Mansoor, G. A., & Massie, B. M. (1999). Left ventricular hypertrophy: A potent cardiovascular risk factor and its relationship to office and ambulatory blood pressure. *Blood Pressure Monitoring*, 4(Suppl 1), S19-S22.

Maringwa, J. T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Ammel, K. V., Teisman, A., & Bijmens, L. (2008a). Application of semiparametric mixed models and simultaneous confidence bands in a cardiovascular safety experiment with longitudinal data. *Journal of*

Biopharmaceutical Statistics, 18(6), 1043-1062. doi:

10.1080/10543400802368881

Maringwa, J. T., Geys, H., Shkedy, Z., Faes, C., Molenberghs, G., Aerts, M., Ammel, K. V., Teisman, A., & Bijmens, L. (2008b). Analysis of cross-over designs with serial correlation within periods using semi-parametric mixed models. *Statistics in Medicine*, 27(28), 6009-6033. doi: 10.1002/sim.3363

Ngo, L., & Wand, M. P. (2004). Smoothing with Mixed Model Software. *Journal of Statistical Software*, 9(1), 1-54. Retrieved from <http://ro.uow.edu.au/eispapers/2649/>

Park, T., & Lee, Y. J. (2002). Covariance models for nested repeated measures data: Analysis of ovarian steroid secretion data. *Statistics in Medicine*, 21(1), 143-164. doi: 10.1002/sim.949

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592. doi: 10.1093/biomet/63.3.581

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric regression*. Cambridge series in statistical and probabilistic mathematics. Cambridge; NY: Cambridge University Press.

Searle, S. R. (2006). *Matrix algebra useful for statistics*. Wiley series in probability and statistics. Hoboken, NJ: Wiley-Interscience.

Shou, H., Zipunniokov, V., Crainiceanu, C. M., & Greven, S. (2014). Structured functional principal component analysis. *Biometrics*, 71(1), 247-257. doi: 10.1111/biom.12236

Silverman, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *The Annals of Statistics*, 24(1), 1-24. doi: 10.1214/aos/1033066196

Snijders, T. (1996). Analysis of longitudinal data using the hierarchical linear model. *Quality and Quantity*, 30(4), 405-426. doi: 10.1007/BF00170145

Szczesniak, R. D., McPhail, G. L., Duan, L. L., Macaluso, M., Amin, R. S., & Clancy, J. P. (2013). A semiparametric approach to estimate rapid lung function decline in cystic fibrosis. *Annals of Epidemiology*, 23(12), 771-777. doi: 10.1016/j.annepidem.2013.08.009

VanDyke, R. D., Ren, Y., Sucharew, H. J., Miodovnik, M., Rosenn, B., & Khoury, J. C. (2012). Characterizing maternal glycemic control: a more informative approach using semiparametric regression. *The Journal of Maternal-Fetal & Neonatal Medicine*, 25(1), 15-19. doi: 10.3109/14767058.2012.626922

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. Springer series in Statistics. New York, NY: Springer.

Wackernagel, H. (2003). *Multivariate geostatistics: an introduction with applications* (3rd ed.). Berlin, Germany: Springer.

Yu, M., & Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statistica Sinica*, 16(4), 1193-1212. Retrieved from <http://www3.stat.sinica.edu.tw/statistica/J16N4/J16N46/J16N46.html>

Appendix A: Semivariogram Description

Revisiting the exponential covariance model from the Methodology section, the semivariogram is:

$$\gamma(h) = C_n + \sigma_0^2 \left[1 - \exp\left(-\frac{|h|}{a_0}\right) \right], \quad |h| > 0$$

The parameters C_n , $C_n + \sigma_0^2$, and a_0 correspond to geostatistical parameters: nugget, sill, and range. The covariance model with $\text{var}(\varepsilon) = C_n + \sigma_0^2$ is called an exponential model with a nugget effect, whereas the covariance model with $\text{var}(\varepsilon) = \sigma_0^2$ is called no-nugget effect model. In a nugget model, σ_0^2 is the partial sill (see Figure A1 below).

Figure A1. Semivariogram of exponential covariance

Appendix B: SAS Implementation

Covariance models (a-d) are presented for the four distinct variance-covariance matrices discussed in the paper, assuming mean response with Structure (1.3) from Table 1. Model structures (1.1-1.5) from Table 1 of the paper may be obtained using the approach described by Maringwa et al. (2008a) but assuming the selected

NESTED SEMIPARAMETRIC MIXED EFFECTS MODELS

covariance function for the DBP data (exponential covariance with nugget effect). Models implemented in SAS are indexed below as 1.3(a-d).

Table B1. Description of variables used in SAS

Variable	Description
DBP	Response variable, diastolic blood pressure
Visit	Occasion of measurement (either pre- or post-intervention)
Studydnr	Subject id for the study
NTime	Time of DBP measurement since sleep onset (in hours)
Timesq	Squared value of NTime
Gender	Indicator variable for gender (1 = Male, 0 = Female)
Bi_race	Indicator variable for race (1 = White, 0 = Non-white)
BMIZ	Continuous variable representing BMI z-score from CDC
NTimecat	Duplicate variable of NTime created for class statement
Z1-Z15	Columns of Z matrix (quadratic) for smoothing ($K = 15$ knots)

Knots were selected using the algorithm from Ngo and Wand (2004). There were 15 knots, ranging from 2.54 to 22.97 hours since sleep onset.

Model 1.3a: Random intercepts

```
proc mixed method=ml data=dataaw;
class studynr visit ntimecat gender bi_race;
model DBP= visit ntime timesq visit*ntime visit*timesq gender bi_race age
bmiz /solution ddfm=kr;
random Z1-Z&nk /type=toep(1) s;
random studynr studynr*visit studynr*ntimecat;
title 'Random intercepts model';
run;
```

Model 1.3b: Composite covariance

```
proc mixed method=ml data=dataaw;
class studynr visit ntimecat gender bi_race;
model DBP=visit ntime timesq visit*ntime visit*timesq gender bi_race age
bmiz/solution ddfm=kr ;
random Z1-Z&nk /type=toep(1) s;
random intercept/subject=studynr s;
repeated visit ntimecat /subject=studynr type=un@ar(1);
```

```

title 'AR(1)@UN composite covariance';
run;

```

Model 1.3c: Exponential covariance (with nugget)

```

proc mixed method=ml data=dataaw convf;
class studynr visit gender bi_race;
model DBP=visit ntime timesq visit*ntime visit*timesq gender bi_race age
bmiz/ddfm=kr s;
random Z1-Z&nk /type=toep(1) s;
random intercept/subject=studynr s;
repeated/subject=studynr type=sp(exp) (ntime_all) local;
title 'SP(EXP) Covariance (with nugget)';
run;

```

Model 1.3d: Exponential covariance (without nugget)

```

proc mixed method=ml data=dataaw convf;
class studynr visit gender bi_race;
model DBP=visit ntime timesq visit*ntime visit*timesq gender bi_race age
bmiz/ddfm=kr solution;
random Z1-Z&nk /type=toep(1) s;
random intercept/subject=studynr s;
repeated/subject=studynr type=sp(exp) (ntime_all);
title 'SP(EXP) Covariance (without nugget)';
run;

```

Bayesian Estimation of the Size of a Street-Dwelling Homeless Population

Lawrence C. McCandless

Simon Fraser University
Burnaby, British Columbia

Michelle L. Patterson

Simon Fraser University
Burnaby, British Columbia

Lauren B. Currie

Simon Fraser University
Burnaby, British Columbia

Akm Moniruzzaman

Simon Fraser University
Burnaby, British Columbia

Julian M. Somers

Simon Fraser University
Burnaby, British Columbia

A novel Bayesian technique is proposed to calculate 95% interval estimates for the size of the homeless population in the city of Edmonton using plant-capture data from Toronto, Canada. The probabilities of capture in Edmonton and Toronto are modeled as exchangeable in a hierarchical Bayesian model, and Markov chain Monte Carlo is used to sample from the posterior distribution. Guidelines are recommended for applying the method to assess the accuracy of homeless counts in other cities.

Keywords: Bayesian statistics, capture-recapture studies, Markov chain Monte Carlo, homelessness

---

Introduction

Estimating the size of street-dwelling homeless populations is important for city planning. However, it is a daunting task that is fraught with methodological and statistical challenges. One strategy is to use a homeless count with the help of volunteers. These volunteers serve as census takers, and their job is to walk throughout the city on predetermined walking routes and interview and count homeless people. For example, in the city of Edmonton, Canada, homelessness counts are conducted every 2 years during a single day in October. [Table 1](#) describes the eight consecutive homeless counts in Edmonton between 1999 and 2012 ([Homeward Trust Edmonton, 2012](#)). [Figure 1a](#) plots the total number of homeless people that were counted during each year.

An astonishing fact about homelessness counts is that interval estimates (e.g. Bayesian 95% credible intervals (CIs)) for the true population size are rarely

---

Dr. McCandless is an Associate Professor in the Faculty of Health Sciences. Email him at: lmccandl@sfu.ca.

provided. For example, the most recent 2012 Edmonton homeless count identified a total of 1070 street-dwelling homeless individuals (see [Table 1](#)). However, no interval estimate was provided. Furthermore, homeless counts are known to be notoriously inaccurate because they underestimate the population size ([Hopper, Shinn, Laska, Meisner, & Wanderling, 2008](#); US Department of Housing and Urban Development, 2008). Homeless people can remain hidden and out of sight. The volunteers can make errors in judgement in determining who is homeless. The street count walking routes may not be sufficiently comprehensive and the number of volunteers may be too few. Variation in counts may also be related to the experience of volunteers and how they are trained. Thus plotted curve in [Figure 1a](#) should be interpreted with extreme scepticism because there is no uncertainty assessment, and it is difficult to judge the accuracy of the estimation.

An important strategy for counting homeless people is to use plant-capture studies ([Schwarz & Seber, 1999](#); [Laska & Meisner, 1993](#); [Martin, Laska, Hopper, Meisner, & Wanderling, 1997](#); [Goudie, Jupp, & Ashbridge, 2007](#); [Hopper et al, 2008](#)). It is a variation of capture-recapture that requires only a single capture. Fake homeless individuals called plants are placed at random locations across the city. The plants are trained to dress and behave in a manner that does not draw attention to themselves so they can blend in with the homeless population. They are assumed to be indistinguishable from other homeless individuals, so that their probability of capture is the same. After the homeless count is complete, the proportion of plants that were counted is examined, and these data are used to estimate the size of the entire homeless population. The plant-capture design is recommended by the United States Department of Housing and Urban Development (2008), which develops guidelines for counting homeless people in American cities.

The validity of the plant-capture methodology depends on several assumptions, and these are reviewed by [Laska and Meisner \(1993\)](#) and [Martin et al. \(1997\)](#). A stable, closed population of individuals is required, with no entry or exits. In practice, this is achieved by conducting the homeless count over a short period of time. Plants should have the same probability of capture as other homeless individuals, and, in particular, the presence of plants should not affect the probability of capture. Plant-capture studies also depend on the accuracy of the data collection. The volunteers must respect the study protocol regarding whom to approach and how to conduct the interview to ascertain homeless status. They should have access to all parts of the street walking routes and a clear understanding of the geography of the city and time restrictions. See [Martin et al. \(1997\)](#) for a review of the assumptions for homeless street counts.

BAYESIAN ESTIMATION OF THE SIZE OF A HOMELESS POPULATION

Table 1. Description of homeless counts in Edmonton and Toronto

	Edmonton in 1999, 2000, ..., 2012,	Toronto in 2006, 2009, 2013
# of street-dwelling homeless who were counted	1070 in 2012, 1533 in 2010; 1862 in 2008; 1774 in 2006; 1452 in 2004; 1213 in 2002; 650 in 2000; 611 in 1999	447 in 2013; 362 in 2009; 735 in 2006
Definition of homelessness	Asking individuals the question: Do you have a permanent residence to return to tonight?"	Any individual sleeping outdoors on the night of the survey
Description	A street count that involved approaching individuals along predetermined walking routes where homeless are known to congregate.	An outdoor survey where teams were instructed to stop everyone they encountered to ask screening questions that establish housing status.
Date, time, temperature and weather conditions	2012: October 16, 05:00 to 22:00, 11.5C, Clear skies; 2010: October 5, 05:00 to 22:00, 10.5C, Clear skies; 2008: October 21, 05:00 to 22:00, 6C, Cloudy skies; 2006: 05:00 October 17 to 05:00 October 18, 0.4C, Clear skies; 2004: 04:30 October 19 to 04:30 October 20, 2.5C, Cloudy skies; 2002: 04:30 October 23 to 05:00 October 24, -3.5C, Clear skies; 2000: September 14, 24 hour period, Temperature and weather unknown; 1999: November 17, 24 hour period, Temperature and weather unknown	2013: April 17, 19:00 to 01:00, 7.5C, Rain showers; 2009: April 15, 19:30 to 11:59, 9C, No precipitation; 2006: April 19, 20:30 to 11:59, 13C, No precipitation
# volunteer enumerators	300 in 2012; 300 in 2010; 220 in 2008; 300 in 2006; 157 in 2004; 200 in 2002; 100 in 2000; 100 in 1999	569 in 2013; 458 in 2009; 750 in 2006
Population of city in 2006	739000	2500000
Area of city in 2006	684 km ²	1749 km ²
Plant capture study?	No	Yes

Figure 1. Estimated size of the homeless population in Edmonton

No plant-capture study has ever been done in Edmonton nor is any planned for the future. Homelessness counts are politically contentious, and controversy surrounds the costs and optics of paying individuals to pretend to be homeless. Thus when interpreting [Figure 1a](#), the analyst is left with a basic research question: Is it possible to build interval estimates to quantify uncertainty in the population size? Is there data that allows us to estimate the proportion of homeless people that were counted during each year?

In this article, a novel Bayesian technique is proposed to calculate 95% interval estimates for the size of the homeless population in Edmonton using external data in the form of plant-capture studies from Toronto, Canada. The Bayesian approach is particularly well-suited to settings where multiple sources of information are available ([Spiegelhalter, Best, Carlin, & Van Der Linde, 2002](#); [Gelman et al., 2013](#)). Synthesizing data into a single model allows propagation of evidence and uncertainty about unknown quantities ([Sweeting, De Angelis, Hickman, & Ades, 2008](#)). This approach is an example of

BAYESIAN ESTIMATION OF THE SIZE OF A HOMELESS POPULATION

Table 2. Plant-capture studies for the Toronto homelessness counts in 2006, 2009, and 2013

Year	Region	Plants Deployed	Plants Found	Proportion of Plants Found
2006	Toronto East-York	24	21	88%
	North York	13	7	54%
	Etobicoke	4	4	100%
	Scarborough	8	6	75%
	Total	50	26	52%
2009	Toronto East-York	17	9	53%
	North York	10	4	40%
	Etobicoke	6	5	83%
	Scarborough	12	8	67%
	Total	45	26	58%
2013	Toronto East-York	18	7	39%
	North York	10	7	70%
	Etobicoke	12	7	58%
	Scarborough	10	5	50%
	Total	49	38	78%

Figure 2. (a) Frequency histogram of the 12 proportions in Table 2. (b) Posterior distribution of the eight quantities $p_H = (p_{H1999}, \dots, p_{H2012})$ calculated from the Bayesian analysis versus the Bayesian analysis with nonparametric regression for the population size

multiparameter evidence synthesis, which combines information from different datasets in order to estimate unknown parameters (Ades & Sutton, 2006).

To outline the proposed Bayesian methodology, consider the Toronto homeless data that is presented in Tables 1 and 2. In 2006, 2009 and 2013, homeless counts were conducted in Toronto, and they included plant-capture studies to estimate the probabilities of capture in Toronto (Toronto Shelter, Support and Housing Administration, 2013). Table 1 describes the homeless counts, and Table 2 summarizes the results of the plant-capture studies. Table 2 shows the number of plants that were deployed to each region of Toronto, by year, and it shows the proportion of plants that were captured. Figure 2a gives a histogram of the 12 proportions from Table 2. The proportions are heterogeneous and range from as low as 39% to as high as 100%. The mean is 65% and standard deviation is 19%. The heterogeneity is due to random error from the small number of plants in each region of Toronto, and additionally, due to variation in the probabilities of capture across space and time.

In this investigation, the histogram in Figure 2a is used to construct a prior distribution for the probability of capture for homeless people in Edmonton. Building on the work of Castledine (1981) and George and Robert (1992), the capture probabilities in Edmonton and Toronto are modelled as exchangeable in a hierarchical Bayesian model. They are treated as a random sample from a Beta distribution with unknown hyperparameters (Coull & Agresti, 1999; Pledger, 2005). The prior distribution expresses our initial beliefs about the probabilities of capture. It is updated using plant-capture studies from Toronto in order to obtain the posterior distribution for the unknown model parameters, including the size of the homeless population in Edmonton during each year.

This article describes the first example of a Bayesian analysis of plant-capture data, and it builds on the Bayesian literature for capture-recapture studies (e.g. Castledine, 1981; Smith, 1991; George & Robert, 1992; Fienberg, Johnson, & Junker, 1999; Basu & Ebrahimi, 2001; King & Brooks, 2001; Tardella, 2002; Manrique-Vallier & Fienberg, 2008; Corkrey et al., 2008). This article is organized as follows: First, the authors describe the methodology and modelling assumptions. An important issue in capture-recapture studies is understanding the role of heterogeneity in probability of capture between individuals (Burnham & Overton, 1978; Coull & Agresti, 1999; Link, 2003; Dozario & Royle, 2003; Pledger, 2005; Hwang & Huggins, 2005; Holzmann, Munk, & Zucchini, 2006; Farcomeni & Tardella, 2012), and this is discussed in the Statistical Models and Methods section. Next, the Results section is presented. The authors describe 95% CIs for the size of the homeless population in Edmonton. Further, the results of a simulation are

presented that examines the sensitivity of the choice of prior distribution on the analysis results, including the coverage probability of interval estimates. A limitation of the analysis is that it ignores the fact that the size of the homeless population should change smoothly over time. Accordingly, in the final section of the Results, the authors incorporate a nonparametric regression model for the population size and study how this impacts uncertainty about the probability of capture. The article concludes with the Discussion section, and we provide guidelines for applying the method to assess the accuracy of homeless counts in other cities.

Statistical Models and Methods

Following Laska and Meisner (1993) and Martin et al. (1997), let H_i for $i \in \{1999, 2000, 2002, 2004, 2006, 2008, 2010, 2012\}$ denote the size of the finite population of homeless people in Edmonton during the homeless count in year i . Let n_{Hi} denote the number of homeless people who were counted in year i . Thus $n_{Hi} \leq H_i$. The quantity n_{Hi} is known, whereas H_i is unknown. The objective is to estimate H_i . The values of n_{Hi} are plotted over time in Figure 1a, and they are listed in the first row of Table 1. For example, $n_{H2012} = 1070$. Write $\mathbf{H}$ and $\mathbf{n}_H$ to denote vectors of the quantities H_i and n_{Hi} over i . Following Laska and Meisner (1993) and Martin et al. (1997), we model n_{Hi} using a Binomial distribution

$$n_{H_i} \sim \text{Binomial}(H_i, p_{H_i}) \quad (1)$$

with size H_i and proportion p_{Hi} . Let $\mathbf{p}_H$ denote the vector of p_{Hi} over index i .

The quantity p_{Hi} is defined as the average of the individual-level probabilities of capture among the H_i homeless people in Edmonton during year i . An important issue in the analysis of plant-capture data is understanding the role of heterogeneity in probability of capture between individuals (e.g. Burnham & Overton, 1978; Coull & Agresti, 1999; Link, 2003; Pedger, 2005). To illustrate the idea of heterogeneity, consider a hypothetical finite population of homeless individuals of size N . Suppose that each individual has only one opportunity for capture. Let $X_l = 1$ or 0, for $l = 1$ to N , be an indicator variable that indicates whether the l th individual was captured. Define $n = \sum_{l=1}^N X_l$ as the total number of homeless individuals who were captured. Additionally, let $P(X_l = 1) = p_l$ denote the individual-level probability of capture, so that $X_l \sim \text{Bernoulli}(p_l)$. Further, suppose that the quantities $p_1, \dots, p_N$ are independent and identically distributed with expected value $E[p_l]$.

Then marginally, averaging over the probability distribution of p_l , we have $X_l \sim \text{Bernoulli}(E[p_l])$ and $n \sim \text{Binomial}(N, E[p_l])$.

Consequently, if one assumes that the detection of homeless individuals in Edmonton are treated as independent events, then this implies that the analyst can model the total number of homeless individuals who are counted in each year using (1), which is a binomial distribution with proportion p_{Hi} and no overdispersion. The quantity p_{Hi} depends on the calendar year i because the proportion of the population that is counted can vary from one year to the next. The Edmonton data are unique because each homeless individual has only one opportunity for capture in year i . In contrast, unmodelled heterogeneity in individual-level capture probabilities can greatly affect estimates of population size in capture-recapture studies because the same individual has multiple opportunities for capture (Burnham & Overton, 1978; Coull & Agresti, 1999; Link, 2003; Pledger, 2005). It can overstate precision about the population size (Link, 2003), and it can produce downward bias due to ignoring individuals with lower capture probabilities (Hwang & Huggins, 2005).

From (1), the conditional probability $P(n_{Hi} | H_i, p_{Hi})$ is

$$P(n_{Hi} | H_i, p_{Hi}) = \binom{H_i}{n_{Hi}} p_{Hi}^{n_{Hi}} (1 - p_{Hi})^{H_i - n_{Hi}} \quad (2)$$

The quantity H_i is large. If p_{Hi} is far from zero or one, then we can replace (2) with the normal approximation to the binomial distribution. The quantity n_{Hi} is modelled as normally distributed with mean $H_i p_{Hi}$ and variance $H_i p_{Hi} (1 - p_{Hi})$ which gives

$$P(n_{Hi} | H_i, p_{Hi}) \propto \left\{ 2H_i p_{Hi} (1 - p_{Hi}) \right\}^{-1/2} \exp \left\{ - \frac{(2H_i p_{Hi} (1 - p_{Hi}))^{-1}}{\times (n_{Hi} - H_i p_{Hi})^2} \right\} \quad (3)$$

This Gaussian approximation can be used to accelerate Markov chain Monte Carlo (MCMC) computation.

The objective is to estimate H_i . A Bayesian approach is used to assign a hierarchical prior distribution to the capture probabilities p_{Hi} over i . To illustrate, write the joint probability density of the quantities (n_{Hi}, H_i, p_{Hi}) as

$$P(n_{Hi}, H_i, p_{Hi}) = P(n_{Hi} | H_i, p_{Hi}) \times P(H_i, p_{Hi})$$

Where $P(H_i, p_{Hi})$ is the joint prior distribution for H_i and p_{Hi} . Following George and Robert (1992) and Tardella (2002), the quantities H_i and p_{Hi} are assumed to be marginally independent a priori (i.e. that $P(H_i, p_{Hi}) = P(H_i)P(p_{Hi})$). There is no reason to believe that the probability of capture depends on the size of the homeless population.

To specify a prior $P(H_i)$, this paper builds on the work of George and Robert (1992), who investigate different prior distributions for sample size in capture-recapture studies, including uniform priors. The following prior distribution is assigned

$$P(H_i) \sim \text{Uniform}(n_{Hi}, M = 10000)$$

which is a uniform distribution for H_i that ranges from n_{Hi} to 10000. This prior ensures that H_i cannot be less than n_{Hi} . Additionally, it has upper limit $M = 10000$ to reflect the prior belief that the size of the homeless population cannot be greater than 10000 individuals. It is important that the prior distribution $P(H_i)$ penalize large values H_i . The reason is because during joint estimation of (p_{Hi}, H_i) the MCMC samplers may fail to converge when p_{Hi} and H_i simultaneously tend to zero and infinity, respectively. Other alternative priors for H_i include the Jeffreys prior $P(H_i) \propto 1/H_i$ (Smith, 1991; George & Robert, 1992) or Rissanen's prior (Tardella, 2002).

To formulate a prior for p_{Hi} , plant-capture data from Toronto is incorporated using a Bayesian hierarchical model. The Bayesian approach is well-suited to settings where multiple data sources are available (Spiegelhalter et al., 2002; Gelman et al., 2013). Referring to the data in Table 2, let R_j denote the number of plants that were deployed in region $j \in \{\text{East York in 2006, North York in 2006, Etobicoke in 2006, Scarborough in 2006, East York in 2009, North York in 2009, Etobicoke in 2009, Scarborough in 2009, East York in 2013, North York in 2013, Etobicoke in 2013, Scarborough in 2013}\}$. Similarly, let n_{Rj} denote the corresponding number of plants that were subsequently captured during the Toronto homeless count. So for example, Table 2 illustrates that $R_{\text{Etobicoke in 2009}} = 6$ and $n_{\text{Etobicoke in 2009}} = 5$. A binomial model is assigned to n_{Rj} , which can be written as

$$n_{R_j} \sim \text{Binomial}(R_j, p_{R_j})$$

where p_{Rj} is the capture probability of the R_j plants. The quantity p_{Rj} depends on j to reflect the fact that the probability of capture may vary by calendar year and region. Let $\mathbf{R}$, $\mathbf{n}_R$, and $\mathbf{p}_R$ denote the vector of quantities R_j , n_{Rj} , and p_{Rj} over j .

For the Toronto data, both R_j and n_{Rj} are known, whereas p_{Rj} is unknown. A prior distribution for the unknown capture probabilities p_{Rj} and p_{Hi} is assigned over i and j by modelling the quantities as exchangeable within a hierarchical Bayesian framework. Following Gelman et al. (2013), a common Beta prior distribution is assigned

$$p_{R_j}, p_{H_i} \sim \text{Beta}(\alpha, \beta) \quad (4)$$

for all i, j , with unknown hyperparameters α and β . Beta priors are common in Bayesian analysis of Binomial proportions because they are conditionally conjugate. If the prior distribution for p_{Rj} or p_{Hi} is a Beta, the posterior will also be a Beta. This allows rapid updating of parameters during MCMC computation.

To complete the specification, a prior distribution is required for the unknown hyperparameters α and β . Following Gelman et al. (2013, Section 5.3), the following prior is assigned

Figure 3. Probabilistic graphical model showing the conditional independence structure between data and unknown parameters in Edmonton and Toronto. Square boxes indicate quantities that are fixed and known, circles indicate unknown quantities. Our objective is to estimate $\mathbf{H} = (H_{1999}, \dots, H_{2012})$, the size of the homeless population in Edmonton for each year

$$P(\alpha, \beta) \propto (\alpha + \beta)^{-\frac{5}{2}} \quad (5)$$

which yields a uniform prior on the standard deviation of the Beta distribution in (4).

The practical interpretation of our method is as follows: The collection of unknown probabilities of capture for Toronto and Edmonton is treated as a random sample from a Beta distribution with unknown hyperparameters α and β . The quantities α and β govern the shape of the distribution and, hence, the uncertainty of capture probabilities. Because one can estimate p_{Rj} for all j , this means that one can estimate α and β . Thus the hierarchical model imposes a probability distribution on p_H , which permits estimation of $\mathbf{H}$. Figure 3 presents a probabilistic graphical model showing the conditional independence structure between data and unknown parameters in Edmonton and Toronto.

The full Bayesian model is written as follows: The joint probability density $P(\mathbf{n}_H, \mathbf{H}, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R, \alpha, \beta)$ is given by

$$\begin{aligned} P(\mathbf{n}_H, \mathbf{H}, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R, \alpha, \beta) = & \left[\prod_i P(n_{H_i} | H_i, p_{H_i}) P(H_i) P(p_{H_i} | \alpha, \beta) \right] \\ & \times \left[\prod_j P(n_{R_j} | R_j, p_{R_j}) P(R_j) P(p_{R_j} | \alpha, \beta) \right] \\ & \times P(\alpha, \beta) \end{aligned}$$

The quantities $(\mathbf{n}_R, \mathbf{R}, \mathbf{n}_H)$ are observed, whereas $(\mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta)$ are unknown. The posterior distribution $P(\mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta | \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H)$ obeys the proportionality

$$\begin{aligned} P(\mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta | \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H) \propto & \left[\prod_i P(n_{H_i} | H_i, p_{H_i}) P(H_i) P(p_{H_i} | \alpha, \beta) \right] \\ & \times \left[\prod_j P(n_{R_j} | R_j, p_{R_j}) P(p_{R_j} | \alpha, \beta) \right] \\ & \times P(\alpha, \beta) \end{aligned} \quad (6)$$

To fit the Bayesian model, MCMC is used in order to draw an approximate sample from the posterior distribution in (6). This yields a Markov chain with stationary distribution that is equal to the posterior distribution (Gelman et al., 2013). Using

the MCMC sample the analyst can study the marginal posterior distribution of $\mathbf{H}$, denoted $P(\mathbf{H} \mid \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H)$, in order to estimate the size of the homeless population in Edmonton. Details of the MCMC algorithm are described in the Appendix. In the analyses that follow, the software R is used (R Development Core Team, 2013). Sampler convergence is assessed using multiple chains and diagnostic tools described by Gelman et al. (2013).

Results

Bayesian Estimation of the Size of the Homeless Population in Edmonton

A preliminary analysis is presented for the idealized scenario where the probability of capture is assumed to be exactly equal to 65% for each and every homeless count in Edmonton between 1999 and 2012. Recall from the Introduction that the value 65% is the sample average of the 12 proportions from Toronto listed in Table 2. Thus a naive estimate of $\mathbf{H}$ is obtained by ignoring uncertainty in the capture probabilities p_{Hi} and setting $p_{Hi} = 65\%$ for all i during MCMC computation. When p_{Hi} is fixed and known, then the analyst can sample from the posterior distribution in (6) by updating $\mathbf{H}$ from (A1) and ignoring $\mathbf{p}_H$ and (α, β) altogether.

Figure 1b gives posterior means and 95% highest posterior density CIs for $\mathbf{H}$. Recall that each component of $\mathbf{H}$ is the size of the homeless population in Edmonton during year i . Compared to Figure 1a, the resulting curve is shifted upwards to reflect that only 65% of the population was counted. The interval estimates are very narrow because we have fixed $p_{Hi} = 65\%$.

Next, the full Bayesian analysis is fitted, which samples from the posterior distribution in (6) and estimates all unknown parameters. The results are plotted in Figure 1c, which depict posterior means and 95% CIs for $\mathbf{H}$. The point estimates are similar to those of Figure 1b, however the interval estimates are dramatically wider to reflect the uncertainty about the parameter vector $\mathbf{p}_H$.

To shed further light on the methodology, the solid curves in Figure 2b depict the posterior distribution of each of the eight quantities $\mathbf{p}_H = (p_{H1999}, \dots, p_{H2012})$, which are the average probabilities of capture in Edmonton during each of the eight homeless counts. The eight solid curves lie on top of one another, and they are a Beta approximation to the histogram in Figure 2a. The posterior mean of each quantity is roughly 55%, and the interquartile ranges are from 47% to 64%. Thus Figure 2 illustrates that the Bayesian method is working as expected. The uncertainty about the probabilities of capture in Edmonton translates into a broad

BAYESIAN ESTIMATION OF THE SIZE OF A HOMELESS POPULATION

range of uncertainty about the size of the homeless population, and this stretches the size of the interval estimates.

A Simulation Study to Examine the Sensitivity of the Prior Distribution on Analysis Results

A difficulty with the preceding analysis is that the results depend heavily on the prior distribution for $\mathbf{p}_H$. If the analyst chooses the “right prior” and the assumption of exchangeability between $\mathbf{p}_H$ and $\mathbf{p}_R$ is reasonable, then the interval estimates for the size of the homeless population in Edmonton will be suitably shifted towards the truth. However many things could go wrong. If the prior distribution for $\mathbf{p}_H$ is poorly chosen then the intervals will miss the truth entirely. Do 95% CIs have 95% frequentist coverage probability? To what extent will the coverage probability deteriorate through a careless choice of prior distribution for $\mathbf{p}_H$?

The coverage probability of 95% CIs is examined using a simulation study. In the Edmonton data example, the quantities $\mathbf{n}_H$, $\mathbf{R}$, and $\mathbf{n}_R$ are known. Suppose that $\mathbf{p}_H^*$ and $\mathbf{H}^*$ denote vectors of the true underlying probabilities of capture and true homeless population size for simulation purposes. A simulation is conducted as follows:

Table 3. Simulation study to examine the sensitivity of the prior distribution for $\mathbf{p}_H$ on the analysis results. Cells give the coverage probability of 95% CIs for the size of the homeless population in Edmonton for each year

Simulation #1 where the true capture probabilities are fixed as $\mathbf{p}_H^*$ for each year								
	Coverage probability of 95% CIs							
	1999	2000	2002	2004	2006	2008	2010	2012
Bayesian analysis assuming $p_H = 65\%$ and ignoring uncertainty	94.4%	95.0%	95.5%	95.2%	95.2%	95.0%	95.2%	93.9%
Bayesian analysis with hierarchical prior, which assumes that $\mathbf{p}_H$ and $\mathbf{p}_R$ are exchangeable	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

Simulation #2 where the true capture probabilities are simulated as $\mathbf{p}_H^* \sim \text{Beta}(\alpha^* = 3.37, \beta^* = 1.84)$ for each year								
	Coverage probability of 95% CIs							
	1999	2000	2002	2004	2006	2008	2010	2012
Bayesian analysis assuming $p_H = 65\%$ and ignoring uncertainty	12.1%	10.7%	7.9%	8.7%	6.6%	5.8%	7.3%	9.6%
Bayesian analysis with hierarchical prior, which assumes that $\mathbf{p}_H$ and $\mathbf{p}_R$ are exchangeable	83.4%	87.4%	81.6%	78.3%	80.5%	77.2%	79.3%	81.1%

1. Conduct a full Bayesian analysis of the homelessness data ($\mathbf{nH}$, $\mathbf{R}$, $\mathbf{nR}$) to obtain 95% CIs, denoted I_{Hi} , for the size of the homeless population H_i in each year i .
2. For t from 1 to 1000:
  - a. For Simulation #1: Set the true probabilities of capture as $p_{H_i}^{(t)} = 65\%$ for each i .
  - b. For Simulation #2: Simulate $p_{H_i}^{(t)} \sim \text{Beta}(\alpha^* = 3.37, \beta^* = 1.84)$ for each i , which is a Beta distribution with mean 65% and standard deviation 19%.
  - c. Given $p_{H_i}^{(t)}$ and n_{Hi} , simulate the true homeless population size $H_i^{*(t)}$ from the conditional distribution $P(H_i | p_{H_i}^{(t)} n_{Hi})$ given in (A1) using MCMC.
  - d. Calculate the coverage indicator variable $Q_i^{(t)} = 1_{H_i^{*(t)} \in I_{Hi}}$ for each year i .
3. Calculate the average coverage probability $(1/1000) \sum_{t=1}^{1000} Q_i^{(t)}$ for each year i .

The results are given in Table 3. Simulation #1 considers the scenario where the true probabilities of capture are equal to 65% during each of the Edmonton homeless counts. As expected, the Bayesian analysis that correctly assumes $p_{Hi} = 65\%$ gives 95% CIs that have correct 95% coverage probability. The Bayesian analysis with hierarchical priors is too conservative and the coverage is 100% during each calendar year. In contrast, Simulation #2 describes the more realistic scenario where the true probabilities of capture $\mathbf{pH}^*$ are heterogeneous and sampled from a Beta distribution with mean 65% and a standard deviation 20% (Gelman et al., 2013). Simulation #2 reveals that the hierarchical Bayesian model gives a large improvement in coverage probability compared to interval estimates that ignore uncertainty in the probability of capture.

Increasing Precision Using Bayesian Nonparametric Regression for the Population Size

One problem with Figure 1c is that the population sizes H_i are estimated independently. The inferences for H_i are driven entirely by n_{Hi} and p_{Hi} (see (A1)). But this ignores the reality that the population size should change smoothly over

time. For example, if we know that $H_{2004} = 2000$, then can we not surmise that H_{2002} and H_{2006} are also close to 2000? This modelling information is ignored in Figure 1c. In other words, Figure 1c uses independent priors for each component of $\mathbf{H}$.

To incorporate dependence in the prior for $\mathbf{H}$, a model is required for the way in which the population size changes over time. Natural cubic splines are used (Gelman et al., 2013)

$$H_i \sim N\left\{\sum_{k=1}^3 \varphi_k g_k(i), \sigma^2\right\}$$

with a single knot at i equal to the year 2005, which is the median of the collection of years. The quantities $g_k(i)$ and φ_k are natural cubic spline basis functions and regression coefficients, respectively, and σ^2 is the unknown variance.

A relatively uninformative prior distributions is assigned to the regression parameters. The following prior is given to the coefficients

$$\varphi_1, \varphi_2, \varphi_3 \sim N(0, 10^3)$$

and the variance is given the prior

$$\sigma^2 \sim \text{Inv-}\chi^2(10^{-3}, 10^{-3}) \quad (7)$$

Write $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)$. The posterior distribution becomes

$$\begin{aligned} & P(\mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta, \boldsymbol{\varphi}, \sigma^2 \mid \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H) \\ & \propto \left[\prod_i P(n_{H_i} \mid H_i, p_{H_i}) P(H_i \mid \boldsymbol{\varphi}, \sigma^2) P(p_{H_i} \mid \alpha, \beta) \right] \\ & \times \left[\prod_j P(n_{R_j} \mid H_j, p_{R_j}) P(p_{R_j} \mid \alpha, \beta) \right] \\ & \times P(\alpha, \beta) P(\boldsymbol{\varphi}) P(\sigma^2) \end{aligned} \quad (8)$$

To fit the regression model using MCMC, additional updates of $\boldsymbol{\varphi}$ and σ^2 are required. However, the required conditional distributions are available analytically using Bayesian linear regression. Details are given in the Appendix.

The results of fitting the model are given in Figure 1d. The posterior means of H are smoother than in Figure 1c because they have been shrunk together to fit the nonparametric curve. Interestingly, the interval estimates for H are sharply contracted compared to Figure 1c. When the analyst assumes that the population changes smoothly over time, then this gives more precise estimates of the population size because the regression model stabilizes the predictions.

Using a nonparametric curve to estimate the population size also implies a reduction in uncertainty about the probabilities of capture $\mathbf{p_H}$. This is illustrated in Figure 2b. The dashed curves plots the posterior distribution of each of the eight quantities $\mathbf{p_H} = (p_{H1999}, \dots, p_{H2012})$. The dashed curves are narrower than the solid curves. The locations of the curves are distorted to assist with fitting the nonparametric curve. This means that if the analyst assumes that the population size changes smoothly over time, then this induces a correlation among the p_{Hi} from one year to the next. The analysis with independent priors for $\mathbf{H}$ is too pessimistic about the magnitude of uncertainty about the probabilities of capture.

Discussion

The most recent homeless count in Edmonton occurred on October 17, 2012. A team of 300 volunteers found 1070 homeless people. Based on the Bayesian analysis that incorporates plant-capture data from Toronto, it is estimated that the true size of the homeless population is 2007 individuals with 95% Bayesian credible interval 1137 to 3042 (see Figure 1c). The city of Edmonton hopes to eliminate homelessness over the next decade, and an important question for government policy-makers is to determine whether the size of the homeless population is decreasing over time. The 2012 Edmonton Homeless Count Report states that “Between 2008 and 2012, the unsheltered homeless decreased by 30%” (Homeward Trust Edmonton, 2012). This calculation was based on the number of homeless people who were counted in 2012 (1070 individuals) versus 2010 (1533 individuals) because $(1533-1070)/1533 = 30\%$. The estimation completely ignores uncertainty in the population size.

In contrast, the proposed Bayesian analysis directly contradicts this conclusion in the government report. The posterior mean of the ratio $(H_{2010} - H_{2012})/H_{2010}$, based on the Bayesian nonparametric regression analysis, is equal to 13% with 95% CI -40% to 58%. This implies a mere 13% reduction in the population size between 2010 and 2012, and there is a huge range of uncertainty and the interval estimate covers zero. Thus this analysis highlights the value of Bayesian uncertainty assessments when estimating the size of street-dwelling

homeless populations. The failure to quantify uncertainty using posterior credible intervals can result in erroneous conclusions, which directly impact government policy decisions.

Our analysis depends on the assumptions that underlie plant-capture studies in general. See Laska and Meisner (1993) and Martin et al. (1997) for review. An important issue in the analysis of capture-recapture data is understanding the role of heterogeneity in probability of capture between individuals (Burnham & Overton, 1978; Coull & Agresti, 1999; Link, 2003; Pledger, 2005). In the analysis it is assumed that the homeless detections are independent events. As described the Statistical Models and Methods section, this assumption implies that the total number of homeless individuals who are counted in each year can be modelled using a binomial distribution with no overdispersion (see (1)). However, the assumption neglects the fact that homeless people usually live in groups (Martin et al., 1997). If homeless people aggregate into small groups, then the whole group is either spotted or lost. In principle, one could extend the modelling approach to model dependence in the probabilities of capture. For example, it is possible to model the probabilities of capture using a mixture of Beta distributions (Coull & Agresti, 1999). However, relaxing the independence assumption can cause the model to be nonidentifiable (Link, 2003).

More generally, the proposed Bayesian method can be used to quantify the accuracy of homeless counts in other cities. For example, Hopper et al. (2008) evaluated a plant-capture study of homelessness in New York City in 2005. The authors estimated the proportion of plants who were counted and, additionally, they conducted postcount interviews of homeless individuals to inquire about their whereabouts on enumeration night in order to establish if they were visible. A different example of plant-captures studies of homelessness is described by Martin (1992). In principle, these data could be used to assess the accuracy of homelessness counts in other American cities. Combining data from different cities requires a careful a careful examination of the exchangeability assumption.

References

Ades, A. E. & Sutton, A. J. (2006). Multiparameter evidence synthesis in epidemiology and medical decision-making: current approaches. *Journal of the Royal Statistical Society: Series A*, 169(1), 5-35. doi: [10.1111/j.1467-985X.2005.00377.x](https://doi.org/10.1111/j.1467-985X.2005.00377.x)

- Basu, S. & Ebrahimi, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika*, 88(1), 269-279. doi: [10.1093/biomet/88.1.269](https://doi.org/10.1093/biomet/88.1.269)
- Burnham, K. P. & Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 65(3), 625-633. doi: [10.1093/biomet/65.3.625](https://doi.org/10.1093/biomet/65.3.625)
- Castledine, B. J. (1981). A Bayesian analysis of multiple-recapture sampling for a closed population. *Biometrika*, 68(1), 197-210. doi: [10.1093/biomet/68.1.197](https://doi.org/10.1093/biomet/68.1.197)
- Corkrey, R., Brooks, S., Lusseau, D., Parsons, K., Durban, J. W., Hammond, P. S., & Thompson, P. M. (2008). A Bayesian capture–recapture population model with simultaneous estimation of heterogeneity. *Journal of the American Statistical Association*, 103(483), 948-960. doi: [10.1198/016214507000001256](https://doi.org/10.1198/016214507000001256)
- Coull, B. A. & Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture–recapture studies. *Biometrics*, 55(1), 294-301. doi: [10.1111/j.0006-341X.1999.00294.x](https://doi.org/10.1111/j.0006-341X.1999.00294.x)
- Dorazio, R. M. & Royle A. J. (2003). Mixture models for estimating the size of a closed population when capture rates vary among individuals. *Biometrics*, 59(2), 351-364. doi: [10.1111/1541-0420.00042](https://doi.org/10.1111/1541-0420.00042)
- Farcomeni, A. & Tardella, L. (2012). Identifiability and inferential issues in capture-recapture experiments with heterogeneous detection probabilities. *Electronic Journal of Statistics*, 6, 2602-2626. doi: [10.1214/12-EJS758](https://doi.org/10.1214/12-EJS758)
- Fienberg, S. E., Johnson, M. S., & Junker B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple lists. *Journal of the Royal Statistical Society, Series A*, 162(3), 383-405. doi: [10.1111/1467-985X.00143](https://doi.org/10.1111/1467-985X.00143)
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, A. B., Vehtai, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). New York, NY: Chapman Hall/CRC.
- George, E. I. & Robert, C. P. (1992). Capture-recapture estimation via Gibbs sampling. *Biometrika*, 79(4), 677-683. doi: [10.1093/biomet/79.4.677](https://doi.org/10.1093/biomet/79.4.677)
- Goudie, I. B. J., Jupp, P. E., & Ashbridge, J. (2007). Plant-capture estimation of the size of a homogeneous population. *Biometrika*, 94(1), 243-248. doi: [10.1093/biomet/asm012](https://doi.org/10.1093/biomet/asm012)

BAYESIAN ESTIMATION OF THE SIZE OF A HOMELESS POPULATION

- Holzmann, H., Munk, A., & Zucchini, W. (2006). On identifiability in capture–recapture models. *Biometrics*, 62(3), 934-936. doi: 10.1111/j.1541-0420.2006.00637_1.x
- Hopper, K., Shinn, M., Laska, E. M., Meisner, M., & Wanderling, J. (2008). Estimating numbers of unsheltered homeless people through plant-capture and postcount survey methods. *American Journal of Public Health*, 98(8), 1438-1442. doi: 10.2105/AJPH.2005.083600
- Homeward Trust Edmonton. (2012). 2012 Edmonton Homelessness Count. Retrieved from <http://www.homewardtrust.ca/images/resources/2013-01-22-11-53FINAL%20%202012%20Homeless%20Count.pdf>
- Hwang, W. & Huggins, R. (2005). An examination of the effect of heterogeneity on the estimation of population size using capture-recapture data. *Biometrika*, 92(1), 229-233. doi: 10.1093/biomet/92.1.229
- King, R. & Brooks, S. P. (2001). On the Bayesian analysis of population size. *Biometrika*, 88(2), 317-336. doi: 10.1093/biomet/88.2.317
- Laska, E. M. & Meisner, M. (1993). A plant-capture method for estimating the size of a population from a single sample. *Biometrics*, 49(1), 209-220. doi: 10.2307/2532614
- Link, W. A. (2003). Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, 59(4), 1123-1130. doi: 10.1111/j.0006-341X.2003.00129.x
- Manrique-Vallier, D. & Fienberg, S. E. (2008). Population size estimation using individual level mixture models. *Biometrical Journal*, 50(6) 1051-1063. doi: 10.1002/bimj.200810448
- Martin, E. (1992). Assessment of S-Night street enumeration in the 1990 Census. *Evaluation Review*, 16(4), 418-438. doi: 10.1177/0193841X9201600407
- Martin, E., Laska, E. M., Hopper, K., Meisner, M., & Wanderling, J. (1997). Issues in the use of a plant-capture method for estimating the size of the street dwelling population. *Journal of Official Statistics*, 13(1), 59-74. Retrieved from <http://www.jos.nu/Articles/abstract.asp?article=13159>
- Pledger, S. (2005). The performance of mixture models in heterogeneous closed population capture–recapture. *Biometrics*, 61(3), 868-873. doi: 10.1111/j.1541-020X.2005.00411_1.x
- Schwarz, C. J. & Seber, G. A. F. (1999). Estimating animal abundance: Review III. *Statistical Science*, 14(4), 427-456. Available from <http://www.jstor.org/stable/2676809>

- Smith, P. J. (1991). Bayesian analyses for a multiple capture-recapture model. *Biometrika*, 78(2), 399-407. doi: [10.1093/biomet/78.2.399](https://doi.org/10.1093/biomet/78.2.399)
- Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society: Series B*, 64(4), 583-639. doi: [10.1111/1467-9868.00353](https://doi.org/10.1111/1467-9868.00353)
- Sweeting, M. J., De Angelis, D., Hickman, M., & Ades, A. E. (2008). Estimating hepatitis C prevalence in England and Wales by synthesizing evidence from multiple data sources. Assessing data conflict and model fit. *Biostatistics*, 9(4), 715-734. doi: [10.1093/biostatistics/kxn004](https://doi.org/10.1093/biostatistics/kxn004)
- Tardella, L. (2002). A new Bayesian method for nonparametric capture-recapture models in presence of heterogeneity. *Biometrika*, 89(4), 807-817. doi: [10.1093/biomet/89.4.807](https://doi.org/10.1093/biomet/89.4.807)
- Toronto Shelter, Support and Housing Administration. (2013). *Street needs assessment results*. Retrieved from <http://www.toronto.ca/legdocs/mmis/2013/cd/bgrd/backgroundfile-61365.pdf>
- United States Department of Housing and Urban Development. (2008). *A guide to counting unsheltered homeless people, 2nd Revision*. Retrieved from https://www.hudexchange.info/resources/documents/counting_unsheltered.pdf

Appendix

Bayesian Computation for Estimating the Homeless Population Size

The Metropolis Hastings algorithm is used to sample from the posterior distribution $P(\mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta \mid \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H)$ given in (6) by updating in blocks. This involves updating from the following conditional distributions

$$\begin{aligned} P(\mathbf{H} \mid \mathbf{n}_H, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R, \alpha, \beta) \\ P(\mathbf{p}_H, \mathbf{p}_R \mid \mathbf{n}_H, \mathbf{H}, \mathbf{n}_R, \mathbf{R}, \alpha, \beta) \\ P(\alpha, \beta \mid \mathbf{n}_H, \mathbf{H}, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R) \end{aligned}$$

To update $\mathbf{H}$, we have from (6)

$$\begin{aligned} P(\mathbf{H} \mid \mathbf{n}_H, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R, \alpha, \beta) \\ &= P(\mathbf{H} \mid \mathbf{n}_H, \mathbf{p}_H) \\ &\propto P(\mathbf{n}_H \mid \mathbf{H}, \mathbf{p}_H) P(\mathbf{H}) \\ &\propto \prod_i P(n_{H_i} \mid H_i, p_{H_i}) P(H_i) \\ &\propto 1_{[n_{H_i}, 10000]} \times \prod_i \left\{ 2H_i p_{H_i} (1 - p_{H_i}) \right\}^{-\frac{1}{2}} \\ &\quad \times \exp \left\{ - \left(2H_i p_{H_i} (1 - p_{H_i}) \right)^{-1} (n_{H_i} - H_i p_{H_i})^2 \right\} \end{aligned} \tag{9}$$

where the last line uses the Gaussian approximation to the binomial distribution given in (3). The quantity $\mathbf{H}$ is updated using a random walk Metropolis Hasting step with proposal distribution that is multivariate normal with mean that is a zero vector and variance that is equal to the identity matrix multiplied by a tuning parameter that is set by trial MCMC simulation runs. In principle, updating $\mathbf{H}$ could be improved by using a proposal distribution that approximates a negative binomial distribution (Castledine, 1981).

Updating $\mathbf{p}_H$ and $\mathbf{p}_R$ from $P(\mathbf{p}_H, \mathbf{p}_R \mid \mathbf{n}_H, \mathbf{H}, \mathbf{n}_R, \mathbf{R}, \alpha, \beta)$ is straightforward because the capture probabilities are conditionally conjugate under a Beta prior and Binomial model for $\mathbf{n}_H$ and $\mathbf{n}_R$. For all i and j , we have

$$\begin{aligned}
P(p_{H_i} | \mathbf{n}_H, \mathbf{H}, \mathbf{n}_R, \mathbf{R}, \alpha, \beta) &= P(p_{H_i} | n_{H_i}, H_i, \alpha, \beta) \\
&= \text{Beta} \left\{ \alpha + n_{H_i}, \beta + (H_i - n_{H_i}) \right\} \\
P(p_{R_j} | \mathbf{n}_H, \mathbf{H}, \mathbf{n}_R, \mathbf{R}, \alpha, \beta) &= P(p_{R_j} | n_{R_j}, R_j, \alpha, \beta) \\
&= \text{Beta} \left\{ \alpha + n_{R_j}, \beta + (R_j - n_{R_j}) \right\}
\end{aligned}$$

Hence updating $\mathbf{p}_H$ and $\mathbf{p}_R$ is accomplished by direct simulation from a vector of independent Beta random variables.

To update α and β , note that $P(\alpha, \beta | \mathbf{n}_H, \mathbf{H}, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R) = P(\alpha, \beta | \mathbf{p}_H, \mathbf{p}_R)$. Then

$$\begin{aligned}
P(\alpha, \beta | \mathbf{p}_H, \mathbf{p}_R) &\propto P(\mathbf{p}_H | \alpha, \beta) P(\mathbf{p}_R | \alpha, \beta) P(\alpha, \beta) \\
&\propto \prod_i \left[\frac{\gamma(\alpha + \beta)}{\gamma(\alpha)\gamma(\beta)} p_{H_i}^{\alpha-1} (1 - p_{H_i})^{\beta-1} \right] \\
&\quad \times \prod_j \left[\frac{\gamma(\alpha + \beta)}{\gamma(\alpha)\gamma(\beta)} p_{R_j}^{\alpha-1} (1 - p_{R_j})^{\beta-1} \right] \\
&\quad \times (\alpha + \beta)^{-\frac{5}{2}}
\end{aligned}$$

Given $(\mathbf{p}_H, \mathbf{p}_R)$, the right hand side of this equation can be evaluated as a function of α and β . Updating from $P(\alpha, \beta | \mathbf{n}_H, \mathbf{H}, \mathbf{p}_H, \mathbf{n}_R, \mathbf{R}, \mathbf{p}_R)$ is achieved using a random walk Metropolis Hastings step with proposal distribution that is independent bivariate normal with mean zero and variance that is a tuning parameter set during initial MCMC runs.

Bayesian Computation for the Non-Parametric Regression Analysis

To sample from the posterior distribution in (8), the same MCMC procedure as the one described above is used except with additional updates of $\boldsymbol{\phi}$ and σ^2 . The required conditional distributions for $\boldsymbol{\phi}$ and σ^2 are

$$\begin{aligned}
P(\boldsymbol{\phi} | \mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta, \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H, \sigma^2) \\
P(\sigma^2 | \mathbf{H}, \mathbf{p}_H, \mathbf{p}_R, \alpha, \beta, \mathbf{n}_R, \mathbf{R}, \mathbf{n}_H, \boldsymbol{\phi})
\end{aligned}$$

BAYESIAN ESTIMATION OF THE SIZE OF A HOMELESS POPULATION

Both of these distributions are conditionally conjugate based on the prior distributions in (7), and the analyst can sample from them directly using Bayesian linear regression (Gelman et al., 2013).

Non-Normality Propagation among Latent Variables and Indicators in PLS-SEM Simulations

Ned Kock

Texas A & M International University
Laredo, Texas

Structural equation modeling employing the partial least squares method (PLS-SEM) has been extensively used in business research. Often the use of this method is justified based on claims about its unique performance with small samples and non-normal data, which call for performance analyses. How normal and non-normal data are created for the performance analyses are examined. A method is proposed for the generation of data for exogenous latent variables and errors directly, from which data for endogenous latent variables and indicators are subsequently obtained based on model parameters. The emphasis is on the issue of non-normality propagation among latent variables and indicators, showing that this propagation can be severely impaired if certain steps are not taken. A key step is inducing non-normality in structural and indicator errors, in addition to exogenous latent variables. Illustrations of the method and its steps are provided through simulations based on a simple model of the effect of e-collaboration technology use on job performance.

Keywords: E-Collaboration; Partial Least Squares; Latent Variable; Indicator; Non-Normal Data; Monte Carlo Simulation

---

Introduction

Structural equation modeling (SEM) employing the partial least squares (PLS) method, or PLS-SEM for short, has been extensively used in business research (Hair, Ringle, & Sarstedt, 2011; Kock, 2010; 2014). It has also been increasingly used in a wide variety of fields; some closely related to business, including subfields, and others less so. Examples are information systems (Guo, Yuan, Archer, & Connelly, 2011; Kock & Lynn, 2012), marketing (Biong & Ulvnes, 2011), international business (Ketkar, Kock, Parente, & Vervielle, 2012), nursing

---

Ned Kock is a Distinguished Professor and Chair of the Division of International Business and Technology Studies. Email him at: nedkock@gmail.com.

(Kim et al., 2012), medicine (Berglund, Lytsy & Westerling, 2012), and global environmental change (Brewer, Cinner, Fisher, Green, & Wilson, 2012).

One of the elements that characterize the PLS-SEM method is that it creates latent variables (sometimes referred to as latent “composites”) by means of weighted aggregations of their respective indicators, where the weights are obtained through iterative algorithms (Cirillo & Barroso, 2012; Lohmöller, 1989). The simple model shown in Figure 1 illustrates the main elements of any model used in PLS-SEM.

Figure 1. Structural model with two latent variables

*Notes: latent variables within ovals; loadings next to indicator arrows.

Our simple model follows from past empirical research (Cassivi, Lefebvre, Lefebvre, & Léger, 2004; Chen, Chen, & Capistrano, 2013). It contains two latent variables, e-collaboration technology use (T) and job performance (P), which are measured indirectly through three indicators each. The unit of analysis is assumed to be a team of individuals who collaborate to accomplish work-related tasks in their respective organizations. E-collaboration technology use (T) measures the extent to which a team uses an integrated technology including e-mail and voice conferencing to facilitate the collaborative work of its members. Job performance (P) measures the performance of each team, as perceived by individuals who receive the outputs of the team to perform downstream work-related tasks.

The structural error ε , when properly weighted, accounts for the variance in the latent variable job performance (P) that is not explained by e-collaboration technology use (T). For e-collaboration technology use (T) the indicators are x_{T1} , x_{T2} and x_{T3} . For job performance (P) the indicators are x_{P1} , x_{P2} and x_{P3} . When properly weighted, the uncorrelated indicator errors θ_{T1} , θ_{T2} , θ_{T3} , θ_{P1} , θ_{P2}

and θ_{p3} account for the variances in the indicators that are not explained by their corresponding latent variables.

The indicators store answers to question-statements in a questionnaire. The question-statements are redundant with one another, with respect to each latent variable, and are assumed to “reflect” the latent variable. That is, the indicators are assumed to measure only the latent variable to which they refer. This measurement carries a certain amount of imprecision, which is indicated by the loadings being lower than 1. This implies the existence of measurement error, which would be absent if at least one loading were to be equal to 1.

Because PLS-SEM algorithms are generally claimed to perform particularly well with small samples and non-normal data (Hair et al., 2011), it is necessary to test that claim by comparing the performance of a PLS-SEM algorithm, such as PLS regression (Kock, 2010), in terms of statistical power, against the performance of a non-PLS algorithm. A common choice of “control” non-PLS algorithm is one where indicators are aggregated to generate latent variable scores in a non-weighted fashion; i.e., indicators are aggregated using the same weight.

Performance analyses usually build on Monte Carlo simulations (Robert & Casella, 2005) whereby multiple samples are created and analyzed using the algorithms that are being compared. The samples are created based on true population coefficients. In this case, these are the standardized regression coefficient ($\beta = .3$) and the loadings ($\lambda_{Ti} = \lambda_{Pi} = .7, i = 1 \dots 3$), which are assumed to exist in the population from which the samples are taken. Both the standardized regression coefficient and the loadings are set by the researcher conducting the Monte Carlo simulations.

We address the issue of how one creates normal and non-normal data for such performance analyses. A simple and effective method is proposed for creating data for exogenous latent variables and errors directly, from which data for endogenous latent variables and indicators is subsequently derived. This method is similar to that proposed by Mattson (1997), incorporating elements that arguably make it simpler.

The discussion of the method places emphasis on the issue of non-normality propagation among latent variables and indicators in PLS-SEM simulations, showing that this propagation can be severely impaired if certain steps are not taken. A key step is to induce non-normality in structural and indicator errors, in addition to exogenous latent variables. This is illustrated through Monte Carlo simulations.

A Method for Creating Normal and Non-Normal Data

Several methods exist to create normal and non-normal data for simulations (Headrick, 2010). Power methods relying on polynomial transformations are perhaps the most widely used (Fleishman, 1978; Headrick, 2002). A special case relies on squaring a standardized normal variable X to obtain a non-normal variable X_n as shown in (1) and (2). In these equations $Rndn(N)$ is a function that returns a different normal random variable each time it is invoked, in the form of a vector with N elements, and $Stdz(\cdot)$ is a function that returns a standardized variable.

$$X \leftarrow Stdz(Rndn(N)) \quad (1)$$

$$X_n \leftarrow Stdz(X^2) \quad (2)$$

This method of creating non-normal data has the advantages of introducing enough non-normality to be useful in robustness tests, and at the same time yielding data that follows a χ^2 distribution with 1 degree of freedom. A number of properties are known for this distribution, including probability limit skewness and kurtosis (a.k.a. excess kurtosis) values. These are $\sqrt{8} \cong 2.828$ and 12, respectively, which combined can be seen as indications of severe non-normality.

Figure 2 shows two histograms. The one on the left is for a normally distributed variable X created based on (1) with $N = 1,000$. The one on the right shows a variable X_n that follows a non-normal distribution created based on (2), applied to the normally distributed variable X . Both variables X and X_n are standardized.

Data generated through this method, as well as variations discussed here, is initially standardized. Unstandardization can be easily accomplished by multiplying by σ and adding μ , where σ and μ are the standard deviation and mean of the desired unstandardized distribution, respectively. Rounding to the closest integer within an ordinal scale (e.g., 1...7) yields unstandardized data on a Likert-type scale.

Not only does the non-normal variable X_n present significant positive skewness (i.e., longer tail on the right) and positive kurtosis (i.e., leptokurtosis, or “peakedness”), but it also contains more extreme outliers than X . As noted in other graphs, this is a common feature of non-normal data created through this method. This makes it useful in robustness stress tests; where claimed robustness

in the presence of non-normality is tested under non-normality conditions that are more extreme than usually found in empirical data.

The univariate method described above can be easily extended to the multivariate case. Multiple exogenous latent variables and errors (i.e., error variables) can be created in the same general way, and non-normality can be propagated from exogenous latent variables and structural errors to endogenous latent variables and indicators. This is discussed in the following sections.

Figure 2. Transforming normal into non-normal data

* Notes: both variables X and X_n are standardized; X follows a normal distribution; X_n follows a χ^2 distribution with 1 degree of freedom; X_n was created based on X .

Data with less severe non-normality can be created using the same general method, by increasing the number of degrees of freedom of the χ^2 distribution used. This can be carried out by adding more than one squared standardized normal variable to generate the non-normal variable, as indicated in (3) and (4).

$$X_i \leftarrow Stdz(Rndn(N)) \quad (3)$$

$$X_n \leftarrow Stdz\left(\sum_{i=1}^k X_i^2\right) \quad (4)$$

The number k of standardized normal variables $X_i (i = 1 \dots k)$ used to generate the non-normal variable X_n equals the number of degrees of freedom of the resulting χ^2 distribution. The probability limit skewness and kurtosis of such a distribution are given by $\sqrt{8/k}$ and $12/k$, respectively. Therefore, we can create data with varying degrees of skewness and kurtosis using various values of k through this generalized version of the method.

For example, if $k = 3$ the non-normal variable X_n will have the following probability limit values for skewness and kurtosis: $\sqrt{8/3} \cong 1.633$ and $12 / 3 = 4$, respectively. Data created with these distributional properties could be used in a robustness test for an intermediated condition that could be referred to as one with “moderate” non-normality, and whose results might be contrasted with those for two other conditions: normal, where $Rndn(N)$ would be used with no transformation; and severely non-normal, where a transformation with $k = 1$ would be used.

Creating Normal and Non-Normal Data for Latent Variables

The method is illustrated based on the simple model presented earlier, which contains only two latent variables, and applies to more complex models, with any number of latent variables. In all cases, latent variables and structural errors are created first, and indicators and corresponding errors are created afterwards.

In this model, the normal data for the exogenous latent variable e-collaboration technology use (T) is created according to (5). This is the predictor latent variable in the model. The non-normal data for this same latent variable (T_n) is created according to (6). Analogously, the normal data for the structural error ε is created according to (7). The corresponding non-normal data for the structural error (ε_n) is created according to (8).

$$T \leftarrow Stdz(Rndn(N)) \quad (5)$$

$$T_n \leftarrow Stdz(T^2) \quad (6)$$

$$\varepsilon \leftarrow Stdz(Rndn(N)) \quad (7)$$

$$\varepsilon_n \leftarrow Stdz(\varepsilon^2) \quad (8)$$

Both T and ε have probability limit values of 0 and 0 for skewness and kurtosis, respectively. Conversely, the non-normal variables T_n and ε_n have both probability limit values of $\sqrt{8} \cong 2.828$ and 12 for skewness and kurtosis, respectively. As discussed earlier, these values refer to a χ^2 distribution with 1 degree of freedom.

The normal data for the endogenous latent variable job performance (P) is created according to (9). This is the criterion latent variable in the model. The non-normal data associated with this latent variable can either propagate exclusively from T_n according to (10), or from both T_n and ε_n according to (11). As will become clear, the latter approach, using (11), is the most advisable of the two. In these equations the structural errors are properly weighted (i.e., given the weight $\sqrt{1-\beta^2}$) to account for the variance in P that is not explained by T .

$$P = \beta T + \sqrt{1-\beta^2} \varepsilon \quad (9)$$

$$\dot{P}_n = \beta T_n + \sqrt{1-\beta^2} \varepsilon \quad (10)$$

$$P_n = \beta T_n + \sqrt{1-\beta^2} \varepsilon_n \quad (11)$$

Figure 3 shows data points and regression lines for three samples, where the predictor latent variable is plotted on the horizontal axis and the criterion latent variable on the vertical axis, and in which: (left) both the predictor latent variable, e-collaboration technology use (T), and structural error are normal (ε); (middle) the predictor is non-normal (T_n) but the error is normal (ε); and (right) both the predictor and error are non-normal (T_n and ε_n , respectively). The sample sizes are 1,000 for the three samples. The data was created based on the foregoing equations, with $\beta = .3$ as specified in our model.

At the top of the graphs are the true sample values of the standardized regression coefficients for each case. Their values are relatively stable across graphs, and close or identical to the true population value ($\beta = .3$) implying robustness in the presence of severe non-normality and outliers. The robustness observed is a characteristic of regression methods in general (Haas & Scheff, 1990; Knez & Ready, 1997), and is one of the reasons why PLS-SEM is also a robust method. PLS-SEM builds heavily on regression methods.

NON-NORMALITY PROPAGATION IN PLS-SEM

Figure 3. Normal and non-normal data for latent variables

* Notes: scales are standardized; left - predictor latent variable and error are normal; middle - predictor is non-normal but error is normal; right - predictor and error are non-normal.

It should be emphasized that these standardized regression coefficients are not calculated based on the indicators. They are calculated directly based on the latent variable scores, which are available in the simulation method we describe here. Therefore, these true sample standardized regression coefficients are not distorted by measurement error. This is a type of error discussed earlier, whose existence is implied by the loadings being lower than 1.

As can be inferred from the graphs, when the predictor latent variable is non-normal but the error is normal (middle), the propagation of non-normality from the predictor latent variable T_n to the criterion latent variable job performance $\dot{P}_n$ is severely impaired. In this case, while skewness and kurtosis for T_n are 2.93 and 11.68 respectively, the criterion latent variable $\dot{P}_n$ is essentially normal (skewness = .16, kurtosis = .28).

Using this approach to create non-normal data in Monte Carlo simulations to test a PLS-SEM algorithm would lead to results supporting the conclusion that the algorithm is robust to non-normality when that may not be the case. In other words, if non-normality propagation is severely impaired, robustness tests would be largely meaningless, and may lead to incorrect conclusions.

However, if the approach associated with the graph at the far right is used, where both the predictor and error are non-normal (right), the propagations of non-normality from the predictor latent variable T_n and error ε_n to the criterion latent variable P_n is largely unimpaired. Here the same values of skewness and

kurtosis for T_n lead to 2.80 and 11.27 for P_n , because a large amount of the non-normality comes from the non-normal error ε_n .

Why is the propagation so severely impaired when the predictor latent variable is non-normal (T_n) but the error is normal (ε)? As it will be clear from our discussion of non-normality propagation from latent variables to indicators, the reason is the magnitude of the propagation coefficient that links the latent variables.

In this case, this propagation coefficient is the standardized regression coefficient β , whose value is .3 in the model. This value is small compared with the propagation coefficient for the error ($\sqrt{1-\beta^2} = \sqrt{1-.3^2} \cong .954$). Small propagation coefficients tend to impair non-normality propagation.

Small propagation coefficients are likely to be commonly found in PLS-SEM models, because standardized partial and full regression coefficients tend to be relatively small (or small enough to impair propagation) in models that are free from vertical and lateral collinearity (Kock & Lynn, 2012). The same applies to path models in general, with or without latent variables, and multiple regression models.

Creating Normal and Non-Normal Data for Indicators

Consider the creation of normal and non-normal data for indicators by creating normal and non-normal data for each of the six indicator errors, expressed generally as θ_{Ti} , θ_{Pi} , θ_{Tn} , and θ_{Pn} ($i = 1 \dots 3$).

The normal data for the indicators associated with the exogenous latent variable e-collaboration technology use (T) and the endogenous latent variable job performance (P) are created according to (12) and (13), respectively.

$$x_{Ti} = \lambda_{Ti}T + \sqrt{1-\lambda_{Ti}^2}\theta_{Ti} \quad (12)$$

$$x_{Pi} = \lambda_{Pi}P + \sqrt{1-\lambda_{Pi}^2}\theta_{Pi} \quad (13)$$

Analogously, the non-normal data for the indicators associated with the non-normal versions of the same latent variables, the exogenous latent variable e-collaboration technology use (T_n) and the endogenous latent variable job performance (P_n), are created according to (14) and (15), respectively.

NON-NORMALITY PROPAGATION IN PLS-SEM

$$x_{Ti_n} = \lambda_{Ti} T_n + \sqrt{1 - \lambda_{Ti}^2} \theta_{Ti_n} \quad (14)$$

$$x_{Pi_n} = \lambda_{Pi} P_n + \sqrt{1 - \lambda_{Pi}^2} \theta_{Pi_n} \quad (15)$$

Unlike the structural error weight, used in the creation of the endogenous latent variable, the weights of the indicator errors will tend to have magnitudes that are similar to the magnitudes of the loadings. In some cases, where measurement precision is high (i.e., high loadings), the weights of the indicator errors will be significantly lower than those of the indicator errors.

For example, a loading of .7 will lead to an indicator error weight of $\sqrt{1 - .7^2} \cong .714$, whereas a loading of .9 will lead to an indicator error weight of $\sqrt{1 - .9^2} \cong .436$. In the former case, the degree of non-normality propagation, measured through the corresponding coefficients of propagation (loading of .7 and weight of .714), will be about the same from the latent variable and the indicator error. In the latter case, the degree of non-normality propagation from the latent variable (loading of .9) will be much greater than from the indicator error (weight of .436).

Figure 4. Normal and non-normal data for indicators

* Notes: scales are standardized; latent variable - T ; indicator - x_{Ti} ; left - latent variable and indicator error are normal; middle - latent variable is non-normal but error is normal; right - latent variable and error are non-normal.

Figure 4 shows data points and regression lines for three samples, where the latent variable is plotted on the horizontal axis and the indicator on the vertical axis, and in which: (left) both the latent variable and the indicator error are normal; (middle) the latent variable is non-normal but the indicator error is normal; and (right) both the latent variable and the indicator error are non-normal. As with the graphs for latent variables, the sample sizes here are 1,000 for the three samples. The data were created based on the foregoing equations with the loadings as specified in our model.

Data for only one latent variable and one indicator are used in these graphs. These variables serve as an illustration of what would happen with any pair of latent variable and corresponding indicator in our model. At the top of the graphs are the true sample values of the loadings for each case.

Non-normality propagation is different for the cases in which the latent variable is non-normal but the indicator error is normal (middle) and both the latent variable and the indicator error are non-normal (right). In the former case, skewness and kurtosis for the latent variable are 2.93 and 11.68 respectively, and 1.05 and 2.67 for the indicator. In the latter case, the same values of skewness and kurtosis for the latent variable lead to 2.05 and 5.24 for the indicator. In neither case non-normality propagates fully; both are examples of partial propagation.

These results bring to the fore two interesting characteristics of non-normality propagation. One is that there is always some loss in the propagation among linked variables; be the propagation among latent variables, or among latent variables and indicators. The other interesting characteristic of non-normality propagation is that the magnitude of the loss is strongly dependent on the propagation coefficients (path coefficients, loadings, and error weights), with the loss increasing steeply in response to decreases in those coefficients.

From these results it seems that this problem is more pronounced in the non-normality propagation from latent variables to indicators, as long as non-normal errors are used – otherwise propagation losses are greater among linked latent variables, because path coefficients tend to be generally lower in magnitude than loadings.

It could be argued that this loss in propagation is not a characteristic of the non-normal data creation method used, but stems from assumptions underlying the common factor model (MacCallum & Tucker, 1991). In it, the propagation of variance (and thus non-normality) happens only from latent variables to indicators, via loadings, and not the other way around.

Skewness and kurtosis values are not usually found in empirical data as extreme as those created. In empirical data, non-normality is often found, but of a

less extreme nature. Therefore, it is possible that the loss in non-normality propagation that we see in our analyses reflects a corresponding phenomenon in actual populations.

Monte Carlo Simulation Results

The results of a set of Monte Carlo simulations are shown in [Figure 5](#) where the performance of a relatively new and increasingly popular PLS-SEM algorithm, namely PLS regression ([Kock, 2010](#)), is shown against a control non-PLS algorithm in the context of our simple model. We used parametric path analysis as the control non-PLS algorithm. WarpPLS version 4.0, was used to analyze the data in our Monte Carlo simulations. The focus of our performance analysis is on statistical power, which is the probability of avoiding false negatives. We created and analyzed 500 samples (or replications) with normal and severely non-normal data. The data were created using the method described in the preceding sections, for each of the following sample sizes: 50, 100, 150, and 200.

The p -value calculation method used for PLS regression is the stable method ([Kock, 2013](#)). This heuristic method employs a built-in table of standard errors generated through bootstrapping and jackknifing ([Chiquoine & Hjalmarsson, 2008](#); [Diaconis & Efron, 1983](#); [Efron et al., 2001](#)), but instead of generating resamples it obtains standard errors based on nonlinear fitting using the built-in table. This significantly increases computational efficiency, particularly when large samples are used. In the parametric path analysis algorithm, which is our “control” non-PLS algorithm, indicators are aggregated to generate latent variable scores using the same weight of 1 for all indicators. The p -value calculation method used for parametric path analysis is the “parametric” method ([Kock, 2013](#)). This method calculates standard errors based on a Student’s t -distribution.

Skewness and excess kurtosis were calculated, and normality tested, for all indicators in each of the generated samples. This was done with the goal of ensuring that, with non-normal data, sample non-normality propagation to indicators occurred to the extent that all indicators followed truly non-normal distributions. Two tests of normality were used, each taking as inputs skewness and excess kurtosis values: the classic Jarque-Bera test ([Jarque & Bera, 1980](#); [Bera & Jarque, 1981](#)) and Gel and Gastwirth’s (2008) robust modification of this test. Both tests, when applied to non-normal data, indicated statistically significantly non-normality in all indicators.

Figure 5. Monte Carlo simulation results

* Notes: vertical axis - statistical power values (probabilities of avoiding false negatives); horizontal axis - sample sizes; PLSR = PLS regression; PATH = parametric path analysis.

As we can see from the results, PLS regression performed better in terms of statistical power than parametric path analysis with both normal and non-normal data, particularly so with small sample sizes. For example, PLS regression reached the widely accepted power threshold of .8 (yielding false negatives 20 percent of the time) with a sample size of approximately 75 with normal data, and with a slightly greater sample size with non-normal data.

Overall both algorithms suffered small performance losses with non-normal data, compared with their performance with normal data. The fact that those losses were small suggests that both algorithms are fairly robust to deviations from normality. This is not surprising because regression techniques in general and related *p*-value calculation methods are generally believed to be remarkably robust to deviations from normality (Haas & Scheff, 1990; Knez & Ready, 1997). PLS-SEM builds heavily on those techniques and methods.

As a side note, we should clarify that the PLS regression algorithm is referred to as “new” in the context of PLS-SEM because it has been more commonly used in the past in chemometrics applications (Wold et al., 2001) not involving PLS-SEM per se. The use of this algorithm in PLS-SEM is growing. It appears to offer some advantages over other PLS algorithms. One of the advantages is the de-coupling of the estimation of coefficients for the structural and measurement models (Kock, 2010), reducing the likelihood of capitalization on error. The advantages tend to become particularly clear when PLS regression is compared with the more widely used PLS mode A (Lohmöller, 1989) in PLS-SEM applications.

Conclusion

A simple and effective method was proposed for the creation of non-normal data that follows a χ^2 distribution with 1 degree of freedom. This gives access to a number of properties, as this is a well known distribution, including probability limit skewness and kurtosis (a.k.a. excess kurtosis) values. These are $\sqrt{8} \cong 2.828$ and 12, respectively, which reflect severe non-normality and are thus useful in robustness tests. It was shown how less severely non-normal data can be generated using the same general approach, by increasing the degrees of freedom of the χ^2 distribution used.

It was shown that proper propagation of non-normality requires the use of non-normal latent variables and errors, which can be created through the same χ^2 distribution approach. It was demonstrated that propagation of non-normality is severely impaired when propagating non-normal latent variables are used in combination with normal errors, and thus that it is important to use errors that are also non-normal. This applies to both structural errors and indicator errors.

Simulation researchers may be tempted to rescale the indicators directly to obtain non-normal data for use in PLS-SEM and other SEM simulations, since the indicators form the “raw material” that is used to compare different SEM techniques. The problem with this approach is that it removes the interdependence between latent variables and indicators, which in turn prevents true sample analyses and comparisons.

The method discussed here generates data for latent variables and errors directly, and then for indicators, preserving that interdependence. It gives full control of the samples, and the ability to calculate a variety of true sample coefficients that are not available from the specified true population model. In fact, this method permits creation of very large samples (e.g., with $N = 10^6$), from which various traits of the population can be ascertained. In samples this large sampling error is very small, and thus coefficients tend to be very similar to those found in the population from which samples are taken. Although the parameterized population model used to create data in simulations allows the true population path coefficients and loadings to be known, it does not inform the shape of the relationship between loadings and weights or the degree of collinearity among latent variables.

The former, the shape of the relationship between loadings and weights, could help us develop better PLS algorithms (Kock, 2010), with unbiased loadings and weights (Cassel et al., 1999). The latter, the degree of collinearity

among latent variables, could help understand the impact that PLS algorithms have on full collinearity variance inflation factors (Kock & Lynn, 2012).

References

- Bera, A. K., & Jarque, C. M. (1981). Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Economics Letters*, 7(4), 313-318. doi:[10.1016/0165-1765\(81\)90035-5](https://doi.org/10.1016/0165-1765(81)90035-5)
- Berglund, E., Lytsy, P., & Westerling, R. (2013). Adherence to and beliefs in lipid-lowering medical treatments: A structural equation modeling approach including the necessity-concern framework. *Patient Education and Counseling*, 91(1), 105-112. doi:[10.1016/j.pec.2012.11.001](https://doi.org/10.1016/j.pec.2012.11.001)
- Biong, H., & Ulvnes, A. M. (2011). If the supplier's human capital walks away, where would the customer go? *Journal of Business-to-Business Marketing*, 18(3), 223-252. doi:[10.1080/1051712X.2011.541375](https://doi.org/10.1080/1051712X.2011.541375)
- Brewer, T. D., Cinner, J. E., Fisher, R., Green, A., & Wilson, S. K. (2012). Market access, population density, and socioeconomic development explain diversity and functional group biomass of coral reef fish assemblages. *Global Environmental Change*, 22(2), 399-406. doi:[10.1016/j.gloenvcha.2012.01.006](https://doi.org/10.1016/j.gloenvcha.2012.01.006)
- Cassel, C., Hackl, P., & Westlund, A. H. (1999). Robustness of partial least-squares method for estimating latent variable quality structures. *Journal of Applied Statistics*, 26(4), 435-446. doi:[10.1080/02664769922322](https://doi.org/10.1080/02664769922322)
- Cassivi, L., Lefebvre, E., Lefebvre, L. A., & Léger, P. M. (2004). The impact of e-collaboration tools on firms' performance. *The International Journal of Logistics Management*, 15(1), 91-110. doi:[10.1108/09574090410700257](https://doi.org/10.1108/09574090410700257)
- Chen, J. V., Chen, Y., & Capistrano, E. P. S. (2013). Process quality and collaboration quality on B2B e-commerce. *Industrial Management & Data Systems*, 113(6), 908-926. doi:[10.1108/IMDS-10-2012-0368](https://doi.org/10.1108/IMDS-10-2012-0368)
- Chiquoine, B., & Hjalmarsson, E. (2008). *Jackknifing stock return predictions*. Washington, DC: Federal Reserve Board.
- Cirillo, M. A., & Barroso, L. P. (2012). Robust regression estimates in the prediction of latent variables in structural equation models. *Journal of Modern Applied Statistical Methods*, 11(1), 42-53. Available at: <http://digitalcommons.wayne.edu/jmasm/vol11/iss1/4>
- Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, 249(1), 116-130.

- Efron, B., Rogosa, D., & Tibshirani, R. (2001). Resampling methods of estimation. In N. J. Smelser, & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13216-13220). New York, NY: Elsevier.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521-532. doi:10.1007/BF02293811
- Gel, Y. R., & Gastwirth, J. L. (2008). A robust modification of the Jarque-Bera test of normality. *Economics Letters*, 99(1), 30-32. doi:10.1016/j.econlet.2007.05.022
- Guo, K. H., Yuan, Y., Archer, N. P., & Connelly, C. E. (2011). Understanding nonmalicious security violations in the workplace: A composite behavior model. *Journal of Management Information Systems*, 28(2), 203-236. doi:10.2753/MIS0742-1222280208
- Haas, C. N., & Scheff, P. A. (1990). Estimation of averages in truncated samples. *Environmental Science & Technology*, 24(6), 912-919. doi:10.1021/es00076a021
- Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *The Journal of Marketing Theory and Practice*, 19(2), 139-152. doi:10.2753/MTP1069-6679190202
- Headrick, T. C. (2002). Fast fifth-order polynomial transforms for generating univariate and multivariate nonnormal distributions. *Computational Statistics and Data Analysis*, 40(4), 685-711. doi:10.1016/S0167-9473(02)00072-5
- Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton, FL: CRC Press.
- Jarque, C. M., & Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, 6(3), 255-259. doi:10.1016/0165-1765(80)90024-5
- Ketkar, S., Kock, N., Parente, R., & Verville, J. (2012). The impact of individualism on buyer-supplier relationship norms, trust and market performance: An analysis of data from Brazil and the U.S.A. *International Business Review*, 21(5), 782-793. doi:10.1016/j.ibusrev.2011.09.003
- Kim, M. J., Park, C. G., Kim, M., Lee, H., Ahn, Y.-H., Kim, E., Yun, S.-N., & Lee, K.-J. (2012). Quality of nursing doctoral education in Korea: Towards policy development. *Journal of Advanced Nursing*, 68(7), 1494-1503. doi:10.1111/j.1365-2648.2011.05885.x

Knez, P. J., & Ready, M. J. (1997). On the robustness of size and book-to-market in cross-sectional regressions. *The Journal of Finance*, 52(4), 1355-1382. doi:[10.2307/2329439](https://doi.org/10.2307/2329439)

Kock, N. (2010). Using WarpPLS in e-collaboration studies: An overview of five main analysis steps. *International Journal of e-Collaboration*, 6(4), 1-11. doi:[10.4018/jec.2010100101](https://doi.org/10.4018/jec.2010100101)

Kock, N. (2013). *WarpPLS 4.0 User Manual*. Laredo, TX: ScriptWarp Systems.

Kock, N. (2014). Advanced mediating effects tests, multi-group analyses, and measurement model assessments in PLS-based SEM. *International Journal of e-Collaboration*, 10(1), 1-13. doi:[10.4018/jec.2014010101](https://doi.org/10.4018/jec.2014010101)

Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems*, 13(7), 546-580.

Lohmöller, J.-B. (1989). *Latent variable path modeling with partial least squares*. Heidelberg, Germany: Physica-Verlag.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common-factor model: Implications for theory and practice. *Psychological Bulletin*, 109(3), 502-511. doi:[10.1037/0033-2909.109.3.502](https://doi.org/10.1037/0033-2909.109.3.502)

Mattson, S. (1997). How to generate non-normal data for simulation of structural equation models. *Multivariate Behavioral Research*, 32(4), 355-373. doi:[10.1207/s15327906mbr3204_3](https://doi.org/10.1207/s15327906mbr3204_3)

Robert, C. P., & Casella, G. (2005). *Monte Carlo statistical methods*. New York, NY: Springer.

Wold, S., Trygg, J., Berglund, A., & Antti, H. (2001). Some recent developments in PLS modeling. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 131-150. doi:[10.1016/S0169-7439\(01\)00156-3](https://doi.org/10.1016/S0169-7439(01)00156-3)

Characterizations of Continuous Distributions by Truncated Moment

M. Ahsanullah
Rider University
Lawrenceville, NJ

M. Shakil
Miami Dade College
Hialeah, FL

B. M. Golam Kibria
Florida International University
Miami, FL

A probability distribution can be characterized through various methods. In this paper, some new characterizations of continuous distribution by truncated moment have been established. We have considered standard normal distribution, Student's t , exponentiated exponential, power function, Pareto, and Weibull distributions and characterized them by truncated moment.

Keywords: Characterization, exponentiated exponential distribution, power function distribution, standard normal distribution, Student's t distribution, Pareto distribution, truncated moment

---

Introduction

Before a particular probability distribution model is applied to fit real world data, it is essential to confirm whether the given probability distribution satisfies the underlying requirements of its characterization. Thus, characterization of a probability distribution plays an important role in statistics and mathematical sciences. A probability distribution can be characterized through various methods, see, for example, Ahsanullah, Kibria, and Shakil (2014), Huang and Su (2012), Nair and Sudheesh (2010), Nanda (2010), Gupta and Ahsanullah (2006), and Su and Huang (2000), among others. In recent years, there has been a great interest in the characterizations of probability distributions by truncated moments. For example, the development of the general theory of the characterizations of probability distributions by truncated moment began with the work of Galambos and Kotz (1978). Further development on the characterizations of probability distributions

---

Dr. Ahsanullah is a Professor of Statistics in the Department of Information Systems and Supply Chain Management. Email him at: ahsan@rider.edu. Dr. Shakil is a Professor of Mathematics in the Department of Arts and Sciences. Email him at: mashakil@mdc.edu. Dr. Kibria is a Professor in the Department of Mathematics and Statistics. Email him at: kibriag@fiu.edu.

by truncated moments continued with the contributions of many authors and researchers, among them Kotz and Shanbhag (1980), Glänzel (1987, 1990), and Glänzel, Telcs, and Schubert (1984), are notable. However, most of these characterizations are based on a simple proportionality between two different moments truncated from the left at the same point. It appears from literature that not much attention has been paid on the characterizations of continuous distributions by using truncated moment. As pointed out by Glänzel (1987) these characterizations may also serve as a basis for parameter estimation. In this paper, some new characterizations of continuous distributions by truncated moment have been established. The organization of this paper is as follows: We will first state the assumptions and establish a lemma which will be needed for the characterizations of continuous distributions by truncated moment. The following section contains our main results for the new characterizations of continuous distributions by truncated moment. Finally, concluding remarks are presented.

A Lemma

This section will state the assumptions and establish a lemma (Lemma 1) which will be useful in proving our main results for the characterizations of continuous distributions by truncated moment.

Assumptions. Let X be a random variable having absolutely continuous (with respect to Lebesgue measure) cumulative distribution function (cdf) $F(x)$ and the probability density function (pdf) $f(x)$. We assume $\alpha = \inf\{x \mid F(x) > 0\}$ and $\beta = \sup\{x \mid F(x) < 1\}$. We define

$$\eta(x) = \frac{f(x)}{F(x)},$$

and $g(x)$ is a differentiable function with respect to x for all real $x \in (\alpha, \beta)$.

Lemma 1. Suppose that X has an absolutely continuous (with respect to Lebesgue measure) cdf $F(x)$, with corresponding pdf $f(x)$, and $E(X \mid X \leq x)$ exists for all real $x \in (\alpha, \beta)$. Then $E(X \mid X \leq x) = g(x)\eta(x)$, where $g(x)$ is a differentiable function and $\eta(x) = \frac{f(x)}{F(x)}$ for all real $x \in (\alpha, \beta)$, if

CHARACTERIZATIONS OF CONTINUOUS DISTRIBUTIONS

$$f(x) = ce^{\int_{\alpha}^x \frac{u-g'(u)}{g(u)} du}$$

where c is determined such that $\int_{\alpha}^{\beta} f(x) dx = 1$.

Note: Since the cdf $F(x)$ is absolutely continuous (with respect to Lebesgue measure), then by Radon-Nikodym Theorem the pdf $f(x)$ exists and hence $\int_{-\infty}^x \frac{u-g'(u)}{g(u)} du$ exists. Also note that, in Lemma 1 above, the left truncated conditional expectation of X considers a product of reverse hazard rate and another function of the truncated point.

Proof of Lemma 1. It is known that

$$\frac{\int_{\alpha}^x u f(u) du}{F(x)} = \frac{g(x) f(x)}{F(x)} .$$

Thus

$$\int_{\alpha}^x u f(u) du = g(x) f(x) .$$

Differentiating both sides of the equation produces the following

$$x f(x) = g'(x) f(x) + g(x) f'(x) .$$

On simplification, one gets

$$\frac{f'(x)}{f(x)} = \frac{x-g'(x)}{g(x)} .$$

Integrating the above equation gives

$$f(x) = ce^{\int_{\alpha}^x \frac{u-g'(u)}{g(u)} du}$$

where c is determined such that $\int_a^b f(x)dx = 1$. This completes the proof of Lemma 1.

Characterizations of some Continuous Distributions by Truncated Moments

Standard Normal Distribution

The characterization of standard normal distribution is provided in Theorem 1 below.

Theorem 1. Suppose that an absolutely continuous (with respect to Lebesgue measure) random variable X has cdf $F(x)$ and pdf $f(x)$ for $-\infty < x < \infty$. We assume that $f'(t)$ and $E(X | X \leq t)$ exist for all t , $-\infty < t < \infty$. Then

$$E(X | X \leq x) = g(x)\eta(x),$$

where

$$\eta(x) = \frac{f(x)}{F(x)}$$

and $g(x) = -1$, if and only if

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \quad \forall x \in (-\infty, \infty),$$

which is the probability density function of the standard normal distribution.

Proof: Suppose

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

CHARACTERIZATIONS OF CONTINUOUS DISTRIBUTIONS

Then it is easily seen that $g(x) = -1$. Consequently, the proof of the “if” part of the Theorem 1 follows from Lemma 1. We will now prove the “only if” condition of the Theorem 1. Suppose that $g(x) = -1$.

Then it easily follows that

$$\frac{f'(x)}{f(x)} = \frac{x - g'(x)}{g(x)} = -x .$$

On integrating the above equation,

$$f(x) = ce^{-\frac{1}{2}x^2} ,$$

where $c = 1/\sqrt{2\pi}$. This completes the proof of Theorem 1.

Student's t Distribution

The characterization of the Student's t is provided in Theorem 2 below.

Theorem 2. Suppose that an absolutely continuous (with respect to Lebesgue measure) random variable X has the cdf $F(x)$ and pdf $f(x)$ for $-\infty < x < \infty$. We assume that $f'(x)$ and $E(X | X \leq x)$ exist for all x , $-\infty < x < \infty$. Then X has the Student's t distribution if and only if

$$E(X | X \leq x) = g(x)\tau(x) ,$$

where

$$g(x) = -\frac{n}{n-1} \left(1 + \frac{x^2}{n} \right), n > 1$$

and

$$\tau(x) = \frac{f(x)}{F(x)} .$$

Proof: Suppose the random variable X has the t distribution with n degrees of freedom. The pdf $f(x)$ of X is

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\left(\frac{n}{2}\right)}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, -\infty < x < \infty.$$

Then it is easily seen that

$$g(x) = \frac{\int_{-\infty}^x u \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\left(\frac{n}{2}\right)}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{u^2}{n}\right)^{-\frac{n+1}{2}} du}{\frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\left(\frac{n}{2}\right)}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}} = -\frac{n}{n-1} \left(1 + \frac{x^2}{n}\right).$$

Consequently, the proof of “if” part of Theorem 2 follows from Lemma 1. Now prove the “only if” condition of the Theorem 2. Suppose that

$$g(x) = -\frac{n}{n-1} \left(1 + \frac{x^2}{n}\right).$$

Then, one easily has

$$g'(x) = -\frac{2x}{n-1}$$

Thus, after simplification, one obtains the following:

$$\frac{x - g'(x)}{g(x)} = \frac{\frac{n+1}{n-1}x}{\frac{n}{n-1}\left(1 + \frac{x^2}{n}\right)} = \frac{\frac{n+1}{2} \frac{2x}{n}}{\left(1 + \frac{x^2}{n}\right)}.$$

Therefore, by Lemma 1, one has

$$f(x) = ce^{\int_{-\infty}^x \frac{\frac{n+1}{2} \frac{2u}{n}}{\left(1 + \frac{u^2}{n}\right)} du} = ce^{-\left(\frac{n+1}{2}\right) \ln\left(1 + \frac{x^2}{n}\right)} = c \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}.$$

Now, using the condition $\int_{-\infty}^{\infty} f(x) dx = 1$, one obtains

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\left(\frac{n}{2}\right)} \Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad -\infty < x < \infty.$$

Note the condition $n \geq 1$ is needed for $E(X|X \leq x)$ to exist. This completes the proof of Theorem 2.

Exponentiated Exponential Distribution

The Characterization of exponentiated exponential distribution is presented in the Theorem 3 below.

Theorem 3. Suppose an absolutely continuous (with respect to Lebesgue measure) random variable X has the cdf $F(x)$ and pdf $f(x)$ for $0 < x < \infty$ such that $f'(x)$ and $E(X|X \leq x)$ exist for all x , $0 < x < \infty$. Then X has the exponentiated exponential distribution

$$f(x) = \alpha \lambda e^{-\lambda x} \left(1 - e^{-\lambda x}\right)^{\alpha-1}, \quad \alpha > 1, x > 0$$

if and only if

$$E(X | X \leq x) = g(x)\eta(x) ,$$

where

$$\eta(x) = \frac{f(x)}{F(x)}$$

and

$$g(x) = \frac{x(1-e^{-\lambda x})^\alpha}{\lambda \alpha e^{-\lambda x} (1-e^{-\lambda x})^{\alpha-1}} - \frac{\int_0^x (1-e^{-\lambda u})^\alpha du}{\lambda \alpha e^{-\lambda x} (1-e^{-\lambda x})^{\alpha-1}}$$

Proof: Suppose

$$f(x) = \alpha \lambda e^{-\lambda x} (1-e^{-\lambda x})^{\alpha-1}, \alpha > 1, x > 0 .$$

Then it is easily seen that

$$g(x) = \frac{x(1-e^{-\lambda x})}{\lambda \alpha e^{-\lambda x}} - \frac{\int_0^x (1-e^{-\lambda u})^\alpha du}{\lambda \alpha e^{-\lambda x} (1-e^{-\lambda x})^{\alpha-1}} .$$

Consequently, the proof of the “if” part of the Theorem 3 follows from Lemma 1. Now prove the “only if” condition of the Theorem 3. Suppose that

$$g(x) = \frac{x(1-e^{-\lambda x})}{\lambda \alpha e^{-\lambda x}} - \frac{\int_0^x (1-e^{-\lambda u})^\alpha du}{\lambda \alpha e^{-\lambda x} (1-e^{-\lambda x})^{\alpha-1}} .$$

Simple differentiation and simplification gives $g'(x) = x + g(x)A(x)$, where

$$A(x) = -\lambda + \frac{(\alpha-1)\lambda e^{-\lambda x}}{1-e^{-\lambda x}}, \alpha > 1 .$$

CHARACTERIZATIONS OF CONTINUOUS DISTRIBUTIONS

Thus

$$\frac{f'(x)}{f(x)} = \frac{x - g'(x)}{g(x)} = -A(x) = -\lambda + \frac{(\alpha - 1)\lambda e^{-\lambda x}}{1 - e^{-\lambda x}} .$$

On integrating the above equation, it follows that

$$f(x) = ce^{\int_0^x A(u) du} .$$

But

$$\begin{aligned} \int_0^x A(u) du &= \int_0^x \left(-\lambda \frac{(\alpha - 1)\lambda e^{-\lambda u}}{1 - e^{-\lambda u}} \right) du \\ &= -\lambda x + (\alpha - 1) \ln(1 - e^{-\lambda x}) \end{aligned}$$

Thus $f(x) = ce^{-\lambda x}(1 - e^{-\lambda x})^{\alpha - 1}$, where

$$\frac{1}{c} = \int_0^\infty e^{-\lambda x} (1 - e^{-\lambda x})^{\alpha - 1} dx = \frac{1}{\alpha \lambda} .$$

This completes the proof of Theorem 3.

Power Function Distribution

The characterization of the power function distribution is provided in Theorem 4 below.

Theorem 4. Suppose an absolutely continuous (with respect to Lebesgue measure) random variable X has the cdf $F(x)$ and pdf $f(x)$ for $0 \leq x < 1$. Assume that $f'(x)$ and $E(X | X \leq x)$ exist for all x , $0 < x < 1$. Then

$$E(X | X \leq x) = g(x)\eta(x) ,$$

where

$$\eta(x) = \frac{f(x)}{F(x)}$$

and

$$g(x) = \frac{x^2}{\alpha + 1},$$

if and only if

$$f(x) = \alpha x^{\alpha-1}, \alpha > 1, 0 < x < 1,$$

which is the pdf of the power function distribution.

Proof: Suppose $f(x) = \alpha x^{\alpha-1}$, $\alpha > 1$, $0 < x < 1$. Then it is easily seen that

$$g(x) = \frac{x^2}{\alpha + 1}.$$

If

$$g(x) = \frac{x^2}{\alpha + 1},$$

then

$$\begin{aligned} x - g'(x) &= \frac{(\alpha - 1)x}{\alpha + 1}, \\ \frac{x - g'(x)}{g(x)} &= \frac{\alpha - 1}{x} \end{aligned}$$

Thus, by Lemma 1,

$$f(x) = ce^{\int_1^x \frac{\alpha-1}{u} du} = cx^{\alpha-1},$$

CHARACTERIZATIONS OF CONTINUOUS DISTRIBUTIONS

where c is a constant. Using the condition $\int_{-\infty}^{\infty} f(x) dx = 1$, we obtain $f(x) = \alpha x^{\alpha-1}$, $\alpha > 1$, $0 < x < 1$. This completes the proof of Theorem 4.

Remark 1. If $\alpha = 1$, then $g(x) = x^2/2$ and one gets a characterization of the uniform distribution in $[0, 1]$.

Pareto Distribution

The characterization of Pareto distribution is provided in Theorem 5 below.

Theorem 5. Suppose the random variable X has an absolutely continuous (with respect to Lebesgue measure) cdf $F(x)$ and pdf $f(x)$. We assume that $F(1) = 0$, $F(x) > 0$ for all $x > 1$, and $E(X)$ exists. Then X has a Pareto distribution if and only if

$$E(X | X \leq x) = g(x) \tau(x) ,$$

where

$$g(x) = \frac{x^{\alpha+1} - x^2}{\alpha - 1}, \quad x > 1, \alpha > 1$$

and

$$\tau(x) = \frac{f(x)}{F(x)}$$

Proof: Suppose the random variable X has the Pareto distribution. The pdf $f(x)$ of X is given by

$$f(x) = \frac{\alpha}{x^{\alpha+1}}, \quad x \geq 1, \alpha > 1 .$$

Then it is easily seen that

$$g(x) = \frac{\int_1^x \frac{\alpha u}{u^{\alpha+1}} du}{\frac{\alpha}{x^{\alpha+1}}} = \frac{x^{\alpha+1} - x^2}{\alpha - 1} .$$

Consequently, the proof of the “if” part of Theorem 5 follows from Lemma 1. Now prove the “only if” condition of the Theorem 5. Suppose that

$$g(x) = \frac{x^{\alpha+1} - x^2}{\alpha - 1}, x > 1, \alpha > 1 .$$

Then it is easy to show that

$$g'(x) = \frac{(\alpha + 1)x^\alpha - 2x}{\alpha - 1}$$

and

$$x - g'(x) = \frac{(\alpha + 1)(x - x^\alpha)}{\alpha - 1} .$$

Consequently,

$$\frac{x - g'(x)}{g(x)} = -\frac{\alpha + 1}{x} .$$

Therefore, by Lemma 1, one obtains

$$f(x) = ce^{-\int_1^x \frac{\alpha+1}{u} du} = \frac{c}{x^{\alpha+1}} .$$

Now, using the condition $\int_{-\infty}^{\infty} f(x) dx = 1$,

$$f(x) = \frac{\alpha}{x^{\alpha+1}}, x \geq 1, \alpha > 1 .$$

CHARACTERIZATIONS OF CONTINUOUS DISTRIBUTIONS

This completes the proof of Theorem 5.

Weibull Distribution

The characterization of Weibull distribution is provided in Theorem 6 below.

Theorem 6. Suppose an absolutely continuous (with respect to Lebesgue measure) random variable X has the cdf $F(x)$ and pdf $f(x)$ for $0 < x < \infty$, and that $f'(x)$ and $E(X | X \leq x)$ exist for all x , $0 < x < \infty$. Then

$$E(X | X \leq x) = g(x)\eta(x)$$

where

$$\eta(x) = \frac{f(x)}{F(x)}$$

and

$$g(x) = \frac{x^{2-\lambda}}{\lambda} + \frac{x^{1-\lambda}}{\lambda} e^{-x^\lambda} h(x) ,$$

with

$$h(x) = \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right) ,$$

where $\gamma(x, n) = \int_0^x u^{n-1} e^{-u} du$, if and only if

$$f(x) = \lambda x^{\lambda-1} e^{-x^\lambda}, \quad 0 < x < \infty, \lambda > 0$$

which is the pdf of the Weibull distribution.

Proof: Note that

$$\begin{aligned} g(x) &= \frac{\int_0^x \lambda u^\lambda e^{-u^\lambda} du}{\lambda x^{\lambda-1} e^{-x^\lambda}} = \frac{x^{2-\lambda}}{\lambda} + \frac{\int_0^x e^{-u^\lambda} du}{\lambda x^{\lambda-1} e^{-x^\lambda}} \\ &= -\frac{x^{2-\lambda}}{\lambda} + \frac{x^{1-\lambda} e^{-x^\lambda}}{\lambda} \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right) \end{aligned}$$

Suppose that

$$g(x) = -\frac{x^{2-\lambda}}{\lambda} + \frac{x^{1-\lambda} e^{-x^\lambda}}{\lambda} \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right).$$

Then

$$\begin{aligned} g'(x) &= -\frac{2-\lambda}{\lambda} x^{1-\lambda} + \frac{x^{1-\lambda}}{\lambda} + \left(\frac{1-\lambda}{\lambda} x^{-\lambda} - \lambda x^{1-\lambda}\right) e^{-x^\lambda} \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right) \\ &= -\frac{1-\lambda}{\lambda} x^{1-\lambda} + \left(\frac{1-\lambda}{\lambda} x^{-\lambda} - \lambda x^{1-\lambda}\right) e^{-x^\lambda} \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right) \end{aligned}$$

Also,

$$\begin{aligned} x - g(x) &= x - \frac{\lambda-1}{\lambda} x^{1-\lambda} - \left(\frac{1-\lambda}{\lambda} x^{-\lambda} - \lambda x^{1-\lambda}\right) e^{-x^\lambda} \gamma\left(x^\lambda, \frac{2\lambda-1}{\lambda}\right) \\ &= \left(\frac{\lambda-1}{x} - \lambda x^{\lambda-1}\right) g(x) \end{aligned}$$

Thus

$$\frac{f'(x)}{f(x)} = \frac{x - g'(x)}{g(x)} = \frac{\lambda-1}{x} - \lambda x^{\lambda-1}.$$

On integrating with respect to x from 0 to x , one obtains $f(x) = \lambda x^{\lambda-1} e^{-x^\lambda}$, where c is constant. On using the boundary conditions $F(0) = 0$ and $F(\infty) = 1$, we have $c = \lambda$ and $F(x) = e^{-x^\lambda}$. This completes the proof of Theorem 6.

Conclusions

Some continuous probability distributions, namely, standard normal, Student's t , exponentiated exponential, power functions, Pareto, and Weibull distributions, are considered. Their corresponding characterizations are provided by truncated moments, which may be useful in applied and physical sciences.

Acknowledgements

The authors would like to thank the referees and the editor for helpful suggestions which improved the quality and presentation of the paper. Author B. M. Golam Kibria dedicates this paper to the person he respects most "Hazrat Gausul Azam Shah Sufi Syed Mahtab Uddin Ahmed (R)" for his love, affection, and showing him the good path in life.

References

- Ahsanullah, M., Kibria, B. M. G., & Shakil, M. (2014). *Normal and Student's t Distributions and Their Applications*, Atlantis Press, France.
- Galambos, J. & Kotz, S. (1978). Characterizations of probability distributions. A unified approach with an emphasis on exponential and related models, *Lecture Notes in Mathematics*, 675. Berlin, Germany: Springer. doi: [10.1007/BFb0069530](https://doi.org/10.1007/BFb0069530)
- Glänzel, W. (1987). A characterization theorem based on truncated moments and its application to some distribution families. In P. Bauer, F. Konecny, & W. Wertz (Eds.), *Mathematical Statistics and Probability Theory* (Vol. B) (75-84). Dordrecht, Netherlands: Reidel. doi: [10.1007/978-94-009-3965-3_8](https://doi.org/10.1007/978-94-009-3965-3_8)
- Glänzel, W. (1990). Some consequences of a characterization theorem based on truncated moments, *Statistics*, 21(4), 613-618. doi: [10.1080/02331889008802273](https://doi.org/10.1080/02331889008802273)
- Glänzel, W., Telcs, A., & Schubert, A. (1984). Characterization by truncated moments and its application to Pearson-type distributions, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 66(2), 173-183. doi: [10.1007/BF00531527](https://doi.org/10.1007/BF00531527)

Gupta, R. C. & Ahsanullah, M. (2006). Some characterization results based on the conditional expectation of truncated order statistics (record values).

Journal of Statistical Theory and Applications, 5, 391-402.

Huang, W. J. & Su, N. C. (2012). Characterizations of distributions based on moments of residual life. *Communications in Statistics - Theory and Methods*, 41(15), 2750-2761. doi: [10.1080/03610926.2011.552827](https://doi.org/10.1080/03610926.2011.552827)

Kotz, S. & Shanbhag, D. N. (1980). Some new approaches to probability distributions. *Advances in Applied Probability*, 12(4), 903-921. doi: [10.2307/1426748](https://doi.org/10.2307/1426748)

Nair, N. U. & Sudheesh, K. K. (2010). Characterization of continuous distributions by properties of conditional variance. *Statistical Methodology*, 7(1), 30-40. doi: [10.1016/j.stamet.2009.08.003](https://doi.org/10.1016/j.stamet.2009.08.003)

Nanda, A. K. (2010). Characterization of distributions through failure rate and mean residual life functions. *Statistics and Probability Letters* 80,(9-10), 752-755. doi: [10.1016/j.spl.2010.01.006](https://doi.org/10.1016/j.spl.2010.01.006)

Su, J. C., and Huang, W. J. (2000). Characterizations based on conditional expectations. *Statistical Papers*, 41(4), 423-435. doi: [10.1007/BF02925761](https://doi.org/10.1007/BF02925761)

The Goldilocks Dilemma: Impacts of Multicollinearity—A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models

Grayson L. Baird
University of Wyoming
Laramie, Wyoming

Stephen L. Bieber
University of Wyoming
Laramie, Wyoming

A common consideration concerning the application of multiple linear regression is the lack of independence among predictors (multicollinearity). The main purpose of this study is to introduce an alternative method of regression originally outlined by Woolf (1951) that eliminates the relatedness between the predictors in a multiple predictor setting.

Keywords: multicollinearity, collinearity, multiple linear regression, ordered variable regression, OVR

Introduction

Social and behavioral scientists often use multiple linear regression (MLR) to answer research questions that involve multiple predictor variables in both experimental and observational research settings. These scientists must consider a host of issues when applying MLR, such as the appropriateness of measurement, sampling, design, and model assumptions. This paper will focus on one commonly encountered problem in the application of MLR: the situation in which the predictor variables are related to one another, a condition generally referred to as multicollinearity.

Although multicollinearity can be defined as a condition in which the predictor variables are correlated with each other to some degree, the literature provides several alternative names and definitions. For instance, Darlington (1968) referred to the relatedness between predictors as intercorrelation. Kutner,

---

Grayson Baird is a research statistician in the Lifespan Biostatistics Core, Rhode Island Hospital. Email him at: grayson_baird@brown.edu. Dr. Bieber is a professor in the Department of Statistics.

Nachtsheim, Neter, and Li (2004) also referred to the relationship between predictors as intercorrelation, but go on to note that extreme intercorrelation is often referred to as multicollinearity. Similarly, both Cohen, Cohen, West, and Aiken (2003), and Nie, Hull, Jenkins, Steinbrenner, and Bent (1975) directly referred to multicollinearity as the situation where two or more predictors are highly intercorrelated with each other. Although Gordon (1968) noted others have used the term multicollinearity, he refers to high correlation among predictors as redundancy, though many texts reserve the term redundancy to denote the squared value of the intercorrelation (see Cohen et al., 2003). Weisberg (2005) refined the concept of multicollinearity by distinguishing different levels of collinearity, where some relatedness between predictors is referred to as approximate collinearity, strong relatedness is referred to as collinearity, and perfect relatedness is referred to as exact collinearity.

Throughout the references above, the terms such as high, extreme, some, and strong are not numerically specified. However, Tabachnick and Fidell (2007) do specifically define thresholds for multicollinearity, indicating that clear multicollinearity exists when predictors correlate above .90, where correlations above .70 may also be suggestive of multicollinearity. For most social and behavioral science researchers, these values are so unattainably high that they could leave the impression that multicollinearity never needs to be considered nor viewed as problematic within their data.

Gordon (1968) noted that "statistics texts focus upon conditions of extremely high correlation because it is at that point that the resulting problems become most nearly statistical ones." (p. 596). Alternatively, Cohen et al. (2003), Kutner et al. (2007), and Weisberg (2005) referred to multicollinearity as a problematic condition, where Nie et al. (1975) refer to multicollinearity as a condition that can cause problems. All the aforementioned authors go on to discuss various problems of multicollinearity in application.

Gordon (1968) observed that discussions of multicollinearity in general are brief in statistical texts and Weisberg (2005) observed that [multi]collinearity itself has no precise definition. Therefore, it appears that multicollinearity does not have a unified definition or meaning and in fact can denote a variety of different concepts in the literature. Given the imprecise nature of the term multicollinearity, the correlation between predictors in this paper will be referred to as simply relatedness. It is hoped that this neutral term expresses the idea of correlation between predictors, without implying unintended connotations such as strength or threshold of correlation, being problematic or not, etc.

THE GOLDBLOCKS DILEMMA

Therefore, the intent of this study aims not at defining multicollinearity, but rather discussing and demonstrating the impacts of related predictors on the MLR model and statistics, for any value of relatedness greater than zero. An alternative to MLR, called ordered variable regression (OVR), will be presented in this paper, which resolves the issue of related predictors entirely by creating and using predictors that are perfectly unrelated.

Relationship between Predictor Variables

The predictor variables in the multiple linear regression (MLR) model can be either independent of each other ($r_{12} = 0$) or correlated to each other ($r_{12} \neq 0$) [for simplicity and without loss of generalizability, only two predictors, X_1 and X_2 , will be considered throughout this paper]. If two predictors are related to each other, then their redundancy (see Cohen et. al. 2003) can be expressed as r_{12}^2 (i.e., the squared value of their correlation; shared variance).

Figures 1 and 2 will be used extensively throughout this article to present the numerous and varied impacts of the relationship between the predictor variables on the response variable (Y). In these Venn diagrams, the area within any circle is equal to 1 (the total variance of any variable = 1.00), thus the partitions of these circles represent proportions of variance (see Kerlinger & Pedhazur, 1973).

Figure 1. Predictors are unrelated

Figure 2. Predictors are related

Figure 1 illustrates the situation when the two predictors are independent (i.e., the circles representing the predictors do not intersect). In this situation, regions 1 and 2 represent the proportions of the response variable Y accounted for by the predictors X_1 and X_2 respectively; specifically the size of region 1 is r_{YX1}^2 and the size of region 2 is r_{YX2}^2 . Region 3 represents that part of Y that can't be predicted by X_1 or X_2 , which will be referred to as the error.

Conversely, Figure 2 illustrates the situation when the two predictors are related to each other (i.e., the circles representing the predictors do intersect). Regions 1 and 2, as they did in Figure 1, represent the unique contributions to the response variable from X_1 and X_2 , respectively. Unlike Figure 1, Figure 2 contains two additional regions, 4 and 5, which reflect the redundancy (r_{12}^2) between the two predictors. The impact of this relationship between the variables X_1 and X_2 complicates the prediction of the response variable by adding a new piece, region 4 (the shared influence of both predictors on Y) to the circle representing Y . In this figure, region 3 is that part of Y which can't be predicted by X_1 and/or X_2 ; once again the error.

Implications and Impacts When the Predictors Are Not Related

Simple Linear Regression

The simple linear regression (SLR) model in which only X_1 is used to predict Y can be expressed as

$$\hat{Y} = b_1 X_1 \quad (1)$$

where b_1 is the least squares estimate of the slope associated with X_1 and is the answer to the research question, "how is X_1 predictive of Y ?" [without loss of generality, it is possible to consider all of the regression models presented in this article from the perspective in which X_1 , X_2 , and Y are standardized. As a consequence and for convenience, the intercept is always 0.] Similarly,

$$\hat{Y} = b_2 X_2 \quad (2)$$

where b_2 is the least squares estimate of the slope associated with X_2 and is the answer to the research question, how is " X_2 predictive of Y ?" Because the variables have all been standardized,

$$b_1 = r_{YX_1} \quad (3)$$

and

THE GOLDBLOCKS DILEMMA

$$b_2 = r_{YX2} \quad (4)$$

There are two related questions to "how is X_i predictive of Y ?" These are "is the relationship between X_i and Y statistically significant," and "how much of Y is predicted by X_i ?" [throughout this paper, the subscript i will be used to designate either X_1 ($i = 1$), X_2 ($i = 2$), or a result derived from X_i]. The first of these two questions is answered by converting the slope into the t -distribution using

$$t_i(n-p-1) = \frac{b_i}{SE(b_i)} = \frac{b_i}{\left[\frac{\sigma_i^2}{(n-1)s^2(X_i)} \right]^{1/2}} \quad (5)$$

where n is the sample size, p is the number of predictors (i.e., $p = 1$), $SE(b_i)$ is the standard error of the slope (b_i), and σ_i^2 is error variance of Y [which is associated with X_i and is estimated using the mean squared error (MSE _{i})], and $s^2(X_i)$ is the variance of the predictor X_i .

The second of these questions is answered using the coefficient of determination or R^2 , which represents the proportion of variance in Y accounted for (explained, predicted) by either

$$R_i^2 = b_i^2 = r_{YX_i}^2 \quad (6)$$

where b_i^2 equals the size of region i . The coefficient of determination can also be calculated through the use of the sums of squares presented in the analysis of variance table. Although unnecessary in this section, it is presented for consistency with subsequent sections of this article. Within the context of SLR, the sums of squares can be partitioned as follows

$$SS_{Total} = SS_{SLR\ Model(i)} + SS_{Error(i)} \quad (7)$$

where SS_{Total} is the total variation found in Y (associated with regions 1, 2, and 3 in Figure 1; as the circles of Figures 1 and 2 have been standardized to variance = 1, the sums of squares are associated with (represented by) the regions in concept, but not equal to them in size.), $SS_{SLR\ Model(i)}$ is the variation in Y associated with predictor variable X_i and $SS_{Error(i)}$ is the variation in Y not associated with the predictor variable X_i . Hence, when $i = 1$ (predicting Y from X_1),

$SS_{SLR \text{ Model}(1)}$ is represented by region 1, and $SS_{Error(1)}$ by regions 2 and 3. Conversely, when $i = 2$ (predicting Y from X_2), $SS_{SLR \text{ Model}(2)}$ is represented by region 2, and $SS_{Error(2)}$ by regions 1 and 3. From this context, the coefficient of determination for SLR models is

$$R_i^2 = \frac{SS_{SLR \text{ Model}(i)}}{SS_{Total}} = \frac{SS_{SLR \text{ Model}(i)}}{SS_{SLR \text{ Model}(i)} + SS_{Error(i)}} \quad (8)$$

The values for R_i^2 as determined by Equations 6 and 8 are identical. Lastly, the significance of R_i^2 can be found by using the Omnibus F -statistic (abbreviated throughout the paper as F),

$$F_i(p, n - p - 1) = \frac{SS_{SLR \text{ Model}(i)} / p}{SS_{Error(i)} / (n - p - 1)} = \frac{SS_{SLR \text{ Model}(i)} / p}{MSE_i} \quad (9)$$

where the mean squared error (MSE_i) is the estimate of the error variance associated with predictor X_i , which was identified as σ_i^2 in Equation 5.

Multiple Linear Regression

Based on the foundational elements for the simple linear regression (SLR) model above, it is possible to develop the multiple linear regression (MLR) model, in which Y is predicted jointly by both X_1 and X_2 . In a parallel form to the preceding section it is possible to start with the fundamental research question, which is "how are X_1 and X_2 jointly predictive of Y ?" The answer to this question is found in the MLR model

$$\hat{Y} = c_1 X_1 + c_2 X_2 \quad (10)$$

where c_1 and c_2 are the least squares estimates of MLR parameters.

Although Equation 10 is considered to be the answer to the question posed, it rests heavily upon how the word jointly is interpreted (this distinction will be considered at length in next section considering the implications and impacts when the predictors are related). In its standard application, MLR produces an additive model (no interaction terms) and thus defines jointly as independent of one another. As a consequence, the coefficient c_1 is actually the answer to the

THE GOLDILOCKS DILEMMA

question, "how is X_1 predictive of Y independent of X_2 ," and c_2 is the answer to the question, "how is X_2 predictive of Y independent of X_1 ?" From this perspective, it can be seen that the coefficients from the MLR answer a similar, yet very distinct question from the context of SLR.

At this point, the most important and logical question is "what is the relationship between c_1 and b_1 , and between c_2 and b_2 ?" Within the context of standardized variables, the MLR coefficients, c_1 and c_2 , can be linked with the bi-variate correlations as follows from Darlington (1968).

$$c_1 = \frac{r_{YX1} - r_{YX2}r_{12}}{1 - r_{12}^2} \quad \text{and} \quad c_2 = \frac{r_{YX2} - r_{YX1}r_{12}}{1 - r_{12}^2} \quad (11)$$

The relationship of Equation 11 with the part correlations (McNemar, 1962), which are also called the semi-partial correlations (Nunnally, 1967), will be discussed at length in the consideration of Equation 27. At this point, the relationship is inconsequential, because $r_{12} = 0$ and as a result Equation 11 reduces to

$$c_i = r_{YXi} = b_i \quad (12)$$

Thus, if the two predictor variables are not related, then the MLR, c_1 and c_2 , are identical to their SLR counterparts, b_1 and b_2 . In addition, the italicized portion of the MLR questions above (independent of) can be deleted and also simplify to their SLR counterparts.

The test of the significance of the regression coefficients c_1 and c_2 is once again found through the t -statistics, which in the context for MLR is

$$t_i(n-p-1) = \frac{c_i}{SE(c_i)} = \frac{c_i}{\left[\left(\frac{\sigma^2}{(n-1)s^2(X_i)} \right) \left(\frac{1}{1-r_{12}^2} \right) \right]^{1/2}} \quad (13)$$

where p is the number of predictors (i.e., $p = 2$). Because $r_{12} = 0$, Equation 13 reduces to

$$t_i(n-p-1) = \frac{c_i}{SE(c_i)} = \frac{c_i}{\left[\frac{\sigma^2}{(n-1)s^2(X_i)} \right]^{1/2}} \quad (14)$$

Although Equation 14 is similar in appearance to Equation 5, they are not identical. The standard errors of the regression coefficients $[SE(c_i)]$ are smaller than the corresponding standard errors $[SE(b_i)]$, because the size of the MSE (σ^2) from the MLR model has been reduced to region 3 only (hence, $\sigma^2 < \sigma_1^2$ and $\sigma^2 < \sigma_2^2$). Therefore, the value of the t -statistics from the MLR model will be larger than in the SLR models; however, they will not necessarily result in smaller p -values given that the degrees of freedom have been reduced by one.

As in the previous section, the MLR answer to the question, "how much of Y is predicted by X_1 and X_2 ," is found using the coefficient of determination. As presented in Darlington (1968), the coefficient of determination within the context of two predictor variables is

$$R^2 = c_1^2 + c_2^2 + 2c_1c_2r_{12} \quad (15)$$

using $r_{12} = 0$ and the result of Equation 12

$$R^2 = b_1^2 + b_2^2 \quad (16)$$

Thus, the coefficient of determination from the multiple regression reduces to the sum of the coefficients of determination from the two separate simple regressions, see Equation 6.

From the context of the partitioning of the sums of squares,

$$SS_{Total} = SS_{MLR Model} + SS_{Error} \quad (17)$$

where

$$SS_{MLR Model} = SS(X_1|X_2) + SS(X_2|X_1) \quad (18)$$

Specifically, $SS(X_1|X_2)$ reflects the amount of variation in Y associated with the first predictor independent of any association with the second predictor

THE GOLDBLOCKS DILEMMA

(represented by region 1 in Figure 1) and $SS(X_2|X_1)$ corresponds with the amount of variation in Y associated with X_2 independent of any association with X_1 (region 2). Given that the predictors are not related, then it can be logically deduced from the results above that $SS(X_1|X_2) = SS_{SLR Model(1)}$, $SS(X_2|X_1) = SS_{SLR Model(2)}$, and

$$SS_{SLR Model} = SS_{SLR Model(1)} + SS_{SLR Model(2)} \quad (19)$$

Hence, the amount of variation in Y accounted for jointly by X_1 and X_2 is simply the sum of their variation from the simple regressions. The simultaneous use of both predictors results in a single model reflecting both predictive regions (1 and 2), while reducing the error to its appropriate minimum (region 3 only). Thus the coefficient of determination becomes

$$\begin{aligned} R^2 &= \frac{SS_{MLR Model}}{SS_{Total}} = \frac{SS_{MLR Model}}{SS_{MLR Model} + SS_{Error}} \\ &= \frac{SS_{SLR Model(1)} + SS_{SLR Model(2)}}{SS_{SLR Model(1)} + SS_{SLR Model(2)} + SS_{Error}} \end{aligned} \quad (20)$$

or the sum of the two coefficients of determination presented in Equation 8.

As with the individual tests of the coefficients, presented in Equation 14, the Omnibus F -statistic is not a simple extension from the SLR results, due to the reduction in the error term and degrees of freedom. The Omnibus F -statistic for the multiple regression is

$$F(p, n-p-1) = \frac{SS_{MLR Model} / p}{SS_{Error} / (n-p-1)} = \frac{(SS_{SLR Model(1)} + SS_{SLR Model(2)}) / p}{MSE} \quad (21)$$

Numerical Example

To illustrate the points made above when considering the SLR and MLR models, and their corresponding results, a numerical example is presented in Table 1 for data in which $r_{12} = .000$. Due to round off errors associated with the standardization of any data set, the actual value of the relatedness of X_1 and X_2 will not be perfectly zero. For these data the relatedness of X_1 and X_2 is 6.00E-18.

The results can be found by using the regression routine in most statistical computer packages. The exception is that a general linear model routine needs to be performed in order to obtain the sums of squares breakdown information specific to each predictor [$SS(X_1|X_2)$ and $SS(X_2|X_1)$] in the MLR context.

In summary, when the predictors are not related, the coefficients produced by the MLR model are identical to the coefficients produced by the SLR models. As a consequence, the R^2 and model sum of squares for the MLR model are the additive composites of the R^2 and model sum of squares produced by the SLR models. Thus, the data of [Table 1](#) confirms the derived results presented in [Equations 12, 16, 19, and 20](#). These results will hold for any data set in which the predictors are unrelated.

Implications and Impacts When the Predictors Are Related

Simple Linear Regression

This section is essentially the duplicate of the simple regression section when the predictors are not related. The primary difference is found in region 4 of [Figure 2](#). What is the impact of this difference on the results presented previously?

The questions of "how is X_i predictive of Y " remain the same and b_i (the estimates of the slopes) are still the answers. However, b_1 is now associated with regions 1 and 4, and b_2 is now associated with regions 2 and 4. Similarly, all of the results presented in [Equations 3 through 9](#) remain the same, but are expanded to include region 4. Hence, any discussion of X_1 now includes both regions 1 and 4, and any discussion of X_2 now includes regions 2 and 4.

It is important to note that even though all of the results are identical, regardless of whether the predictors are related or not, the answers to the fundamental questions, "how is X_i predictive of Y ," are now more complex. The first predictor is no longer solely predictive of Y (represented by region 1), but this prediction is now supplemented by a shared element associated with the second predictor (region 4). The same situation exists when the focus of the SLR is the second predictor. As a consequence, although the fundamental regression questions remain simple, the answers aren't. Unfortunately these two aspects of the predictor variables are fused together in the answers b_i and can't be separated within the context of SLR.

THE GOLDBLOCKS DILEMMA

Table 1. Comparison of Simple and Multiple Regression when the predictors are not related

	SLR(X_1)	SLR(X_2)	MLR	Comments
Coefficient				
X_1	$b_1 = .467$		$c_1 = .467$	$b_1 = c_1 = r_{YX_1}$, Equations 3, 11
X_2		$b_2 = .312$	$c_2 = .312$	$b_2 = c_2 = r_{YX_2}$, Equations 4, 11
Sums of Squares				
X_1	16.115		16.115	MLR result = SLR result for X_1 , Discussion for Equation 19
X_2		7.222	7.222	MLR result = SLR result for X_2 , Discussion for Equation 19
Model	16.115	7.222	23.337	MLR result = sum of the SLR results for X_1 and X_2 , Equation 19
Error	57.884	66.778	50.663	
Total	74.000	74.000	74.000	
R^2 Estimates				
R^2	.218	.097	.315	MLR result = sum of the SLR results for X_1 and X_2 , Equation 16
R^2 from SS			.315	$SS_{MLR\ Model} / SS_{Total} = 23.337 / 74.000 = .315$, Equation 20

*Note: $r_{12} = .000$, $r_{YX_1} = .467$, $r_{YX_2} = .312$, $n = 75$

It is important to note that even though all of the results are identical, regardless of whether the predictors are related or not, the answers to the fundamental questions, "how is X_i predictive of Y ," are now more complex. The first predictor is no longer solely predictive of Y (represented by region 1), but this prediction is now supplemented by a shared element associated with the second predictor (region 4). The same situation exists when the focus of the SLR is the second predictor. As a consequence, although the fundamental regression questions remain simple, the answers aren't. Unfortunately these two aspects of the predictor variables are fused together in the answers b_i and can't be separated within the context of SLR.

Looking at Figure 2 it can be seen that SS_{Total} is now represented by regions 1, 2, 3, and 4. As a result, the SLR for X_1 produces

$$SS_{SLR\ Model(1)} = SS(X_1|X_2) + SS_{Shared} \quad (22)$$

which corresponds with regions 1 and 4, and an $SS_{Error(1)}$ corresponding to regions 2 and 3. [SS_{Shared} will be defined later in Equations 41 and 42.] Similarly, the result of the SLR for X_2 produces

$$SS_{SLR\ Model(2)} = SS(X_2|X_1) + SS_{Shared} \quad (23)$$

which corresponds with regions 2 and 4, and an $SS_{Error(2)}$ corresponding to regions 1 and 3.

Multiple Linear Regression

The previous section with MLR when the predictors were not related began with the logical research question, "how are X_1 and X_2 jointly related to Y ?" However, because the two predictor variables are now related, the definition of the word jointly is much more complicated than in this previous section. In fact, there are now at least three distinct definitions of this word, which each lead to decidedly different conclusions in regard to the regression coefficients, coefficients of determination, sums of squares, and statistical tests.

Definition 1. Jointly is viewed as the composite of the influence of X_1 to Y and the influence of X_2 to Y . This definition reflects jointly as the sum of the two separate SLR questions, "how is X_1 predictive of Y " and "how is X_2 predictive of Y ." The answer to this question is

THE GOLDBLOCKS DILEMMA

$$\hat{Y} = b_1X_1 + b_2X_2 \quad (24)$$

Using Equation 10 and the result of Equation 12, it was found that when $r_{12} = 0$ it is possible for MLR to generate the model presented in Equation 24. However, when $r_{12} \neq 0$, Equation 24 can't be estimated by any single regression model, because b_1 and b_2 must be estimated separately. Thus, Equation 24 should only be considered as a conceptual combination of the two predictors.

From the previous section, the coefficients of determination for these two simple regressions are R_1^2 for X_1 (regions 1 and 4 in Figure 2) and R_2^2 for X_2 (regions 2 and 4 in Figure 2). As a result, if the two were added together to provide a combined estimate, then

$$R^2 = R_1^2 + R_2^2 = b_1^2 + b_2^2 = \text{Region 1} + \text{Region 4} + \text{Region 2} + \text{Region 4} \quad (25)$$

Thus, the combined estimate presented in Equation 25 would double count region 4 and artificially inflate the jointly determined R^2 by the size of region 4. This was not the case for Equation 16, because region 4 didn't exist. Hence, the use of this definition to determine the joint R^2 is accurate only when the predictors aren't related.

In practice, this first definition of jointly would result in answering the multiple regression question from the context of performing two simple regressions and combining their results at the level of discussion rather than at the level of a predictive model. Although the multiple application of SLR in the presence of multiple predictors may be found in the literature, their results should be viewed with considerable caution. As pointed out in the section above, their answers are not as simple as their questions imply (they can't be interpreted independently), and the R^2 from their conceptual combination (jointly determined influence) will increasingly be over estimated as $|r_{12}|$ increases (increasing the size of region 4).

Definition 2. Jointly is viewed as the composite of the influence of X_1 to Y independent of X_2 and the influence of X_2 to Y independent of X_1 . In this context the word jointly reflects a simultaneous relationship and leads directly to the traditional MLR model

$$\hat{Y} = c_1X_1 + c_2X_2 \quad (26)$$

In appearance this is exactly Equation 10. However, is it? As presented in Equation 11, duplicated here, it is known that

$$c_1 = \frac{r_{YX1} - r_{YX2}r_{12}}{1 - r_{12}^2} \quad \text{and} \quad c_2 = \frac{r_{YX2} - r_{YX1}r_{12}}{1 - r_{12}^2} \quad (27)$$

To begin, given that $r_{12} \neq 0$, Equation 27 doesn't simplify as Equation 11 did, and the MLR coefficients (c_i) won't equal their SLR counterparts (b_i). A close inspection of Equation 27 reveals that the MLR coefficients are functions of the part correlations (McNemar, 1962) [although it is common to speak about the multiple regression coefficients as addressing the question of the relationship between a predictor and dependent variable partialling out the influence of other predictors, this process as actually accomplished through the part correlations, not the partial correlations. Symbolically, $r_{YX1.X2}$ refers to the partial correlation and $r_{Y(X1.X2)}$ refers to the part correlation.] The part correlation of X_1 with Y removing the influence of X_2 from Y only (directly represented by region 1 in Figure 2) is

$$r_{Y(X1.X2)} = \frac{r_{YX1} - r_{YX2}r_{12}}{\sqrt{1 - r_{12}^2}} \quad (28)$$

and of X_2 with Y removing the influence of X_1 from Y only (represented by region 2) is

$$r_{Y(X2.X1)} = \frac{r_{YX2} - r_{YX1}r_{12}}{\sqrt{1 - r_{12}^2}} \quad (29)$$

As a consequence, substituting Equations 28 and 29 into Equation 27, the coefficients from the MLR model are

$$c_1 = \frac{r_{Y(X1.X2)}}{\sqrt{1 - r_{12}^2}} \quad \text{and} \quad c_2 = \frac{r_{Y(X2.X1)}}{\sqrt{1 - r_{12}^2}} \quad (30)$$

Thus, although Equation 26 looks very similar to Equation 10, it is dramatically different. This is the first impact of the relatedness of the predictors; the MLR regression coefficients are no longer equal to their SLR counterparts. In MLR the coefficients, through their association with the process of part correlation, have

THE GOLDBLOCKS DILEMMA

had the shared influence (represented by region 4) removed in comparison to the coefficients from SLR. This is the direct result of the additive nature of the MLR model presented in Equation 26.

What additional impact does this second definition of jointly have on the other results in the multiple predictor setting? Within the context of MLR, the test of the significance of the regression coefficients c_1 and c_2 is found through the t -statistic [Equation 13 is duplicated below]

$$t_i(n-p-1) = \frac{c_i}{SE(c_i)} - \frac{c_i}{\left[\left(\frac{\sigma^2}{(n-1)s^2(X_i)} \right) \left(\frac{1}{1-r_{12}^2} \right) \right]^{1/2}} \quad (31)$$

Unlike Equation 13, Equation 31 doesn't reduce to Equation 14 because $r_{12} \neq 0$. As a note, $1/(1-r_{12}^2)$ of Equation 31 is commonly referred to as the Variance Inflation Factor (VIF). Hence, the impact of the relatedness between the two predictors is the inflation of SE (because the VIF must be greater than 1), which results in a decrease in the magnitude (and thus significance) of the t -statistic. The second impact of the relatedness of the predictors is that their independent contributions to predicting Y are less statistically significant.

What is the impact on the coefficient of determination? From Figure 2, it can be seen that R^2 should be the combined influence from X_1 and X_2 independently (region 1 and 2), and the shared influence of X_1 and X_2 (region 4), such that

$$R^2 = \text{region 1} + \text{region 2} + \text{region 4} \quad (32)$$

When $r_{12} = 0$, it is easy to relate the regions of Figure 1 with the components of the R^2 ; as found in Equation 6. However, now that $r_{12} \neq 0$, how do the results from the MLR model correspond with the components of R^2 ? Using Equations 28 and 29 along with the research question posed by the definition of jointly as simultaneously, it can be seen that the MLR model, found in Equation 26, produces

$$\text{region 1} = r_{Y(X_1.X_2)}^2 \quad \text{and} \quad \text{region 2} = r_{Y(X_2.X_1)}^2 \quad (33)$$

such that

$$R^2_{MLR\ Model} = \text{region 1} + \text{region 2} = r^2_{Y(X1..X2)} + r^2_{Y(X2..X1)} \quad (34)$$

and

$$R^2 = R^2_{MLR\ Model} + R^2_{Shared} \quad (35)$$

where R^2_{Shared} equals the size of region 4. Substituting the results of Equation 30 into Equation 33 produces the association between the regions of Figure 2 and the MLR coefficients as

$$\text{region 1} = (1 - r_{12}^2)c_1^2 \quad \text{and} \quad \text{region 2} = (1 - r_{12}^2)c_2^2 \quad (36)$$

Recalling that R^2 for the MLR equals Equation 15, the size of region 4 can be established in terms of the part correlations as

$$\text{region 4} = R^2_{Shared} = \frac{r_{12}}{1 - r_{12}^2} \left[r_{12}r_{Y(X1..X2)}^2 + r_{12}r_{Y(X2..X1)}^2 + 2r_{Y(X1..X2)}r_{Y(X2..X1)} \right] \quad (37)$$

and in terms of the MLR coefficients as

$$\text{region 4} = R^2_{Shared} = r_{12} \left[r_{12}c_1^2 + r_{12}c_2^2 + 2c_1c_2 \right] \quad (38)$$

It can be seen from this discussion that R^2 is actually a combination of two separate and independent pieces; that piece associated with the model ($R^2_{MLR\ Model}$; regions 1 and 2) and that piece associated with the shared influence (R^2_{Shared} ; region 4). The third impact of the relatedness of the predictors is that the R^2 is unequal to $R^2_{MLR\ Model}$, being inflated by the size of region 4, unless $r_{12} = 0$.

These results for the R^2 can also be illustrated by examining the sum of squares. The determination of the sums of squares using this second definition of jointly is often referred to as Type III sums of squares, which is presented in Equation 40.

$$SS_{Total} = SS_{MLR\ Model} + SS_{Shared} + SS_{Error} \quad (39)$$

THE GOLDBLOCKS DILEMMA

where

$$SS_{MLR\ Model} = SS(X_1|X_2) + SS(X_2|X_1) \quad (40)$$

and

$$SS_{Shared} = SS(X_1, X_2) = SS_{SLR\ Model(1)} - SS(X_1|X_2) \quad (41)$$

from Equation 22 and

$$SS_{Shared} = SS(X_1, X_2) = SS_{SLR\ Model(2)} - SS(X_2|X_1) \quad (42)$$

from Equation 23. As expressed in Equation 18, $SS(X_1|X_2)$ reflects the amount of variation in Y associated with the first predictor independent of any association with the second predictor (region 1) and $SS(X_2|X_1)$ corresponds with the amount of variation in Y associated with X_2 independent of any association with X_1 (region 2). SS_{Shared} reflects the joint influence of X_1 and X_2 (represented by region 4), and SS_{Error} now correctly corresponds with region 3 only.

In many textbooks and statistical programs, it appears that the SS_{Model} is not calculated directly, but rather determined indirectly through the simple subtraction whereby $SS_{Model} = SS_{Total} - SS_{Error}$. This calculation works perfectly when $r_{12} = 0$, but when $r_{12} \neq 0$ it mistakenly includes the SS_{Shared} in the SS_{Model} and inflates the sums of squares associated with the model, such that

$$SS_{Model} = SS_{Total} - SS_{Error} = SS_{MLR\ Model} + SS_{Shared} \quad (43)$$

This is perhaps best explained and illustrated by Woolf (1951, see p. 113). Therefore, the R^2 can be calculated using the sums of squares as

$$R^2_{MLR\ Model} = \frac{SS_{MLR\ Model}}{SS_{Total}} = \frac{SS(X_1|X_2) + SS(X_2|X_1)}{SS_{Total}} \quad (44)$$

and as

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{SS_{MLR\ Model} + SS_{Shared}}{SS_{Total}} \quad (45)$$

It can be seen in Equation 45 that the R^2 calculated by the simple subtraction method is once again inflated by SS_{Shared} (region 4) in comparison to the $R^2_{MLR Model}$, as presented in Equation 44.

The last impact of the relatedness of the predictors on the MLR results is seen in the determination of the Omnibus F -statistic

$$F(p, n-p-1) = \frac{SS_{Model}/p}{SS_{Error}/(n-p-1)} = \frac{(SS_{MLR Model} + SS_{Shared})/p}{MSE} \quad (46)$$

whose value can be partitioned such that the components of F are equal to the sum of

$$F_{MLR Model} = \frac{SS_{MLR Model}/p}{SS_{Error}/(n-p-1)} = \frac{[SS(X_1|X_2) + SS(X_2|X_1)]/p}{MSE} \quad (47)$$

and

$$F_{Shared} = \frac{SS_{Shared}/p}{SS_{Error}/(n-p-1)} = \frac{SS_{Shared}/p}{MSE} \quad (48)$$

As with R^2 , the use of SS_{Model} results in the inflation of the F by SS_{Shared} (region 4).

In summary, the MLR coefficients are the direct answers to the research questions posed at the beginning of this section (Definition 2 of the word jointly) and the t -statistics provide the appropriate significance tests of these relationships. However, both the coefficient of determination and the Omnibus F -statistic are inflated in relation to the MLR model by a function of the amount of shared variance (region 4). Hence, the MLR model (c_1 and c_2) is not consistent with the commonly reported summary statistics (R^2 and F). These results will be demonstrated in the numerical example section below.

Definition 3. Jointly is viewed as the composite of the influence of X_1 to Y (from Definition 1) and the influence of X_2 to Y independent of X_1 (Definition 2). In this context, the word jointly affects an ordered relationship (note either X_1 or X_2 can be represented in the first question, with the other predictor in the second. For convenience only, X_1 will be used in the first question and X_2 in the second). Together this ordered relationship could be represented in the model

THE GOLDBLOCKS DILEMMA

$$\hat{Y} = b_1 X_1 + c_2 X_2 \quad (49)$$

where b_1 comes from Equation 24 and c_2 comes from Equation 26. This model will be referred to here as Ordered Variable Regression (OVR). [Unlike Definition 1 that was only a conceptual combination of the two predictors, Definition 3 actually leads to a determinable model, which will be presented later in this section.]

Another way of viewing these two influences is from the context of stepwise regression, in which b_1 is the answer to the question, "what does X_1 contribute to Y ," and c_2 is the answer to the question, "what does X_2 contribute to Y beyond what is already being contributed by X_1 ?"

The significance of these two regression coefficients have already been presented in Equation 14 and Equation 31, respectively. The determination of the sums of squares using this third definition of jointly is often referred to as Type I sums of squares, which is presented in Equation 51.

$$SS_{Total} = SS_{OVR Model} + SS_{Error} \quad (50)$$

Specifically,

$$SS_{OVR Model} = SS_1 + SS_2 \quad (51)$$

where

$$SS_1 = SS(X_1 | X_2) + SS_{Shared} \quad (52)$$

is consistent with Equation 22, which corresponds with regions 1 and 4, and

$$SS_2 = SS(X_2 | X_1) \quad (53)$$

corresponds with region 2, hence

$$SS_{OVR Model} = SS(X_1 | X_2) + SS(X_2 | X_1) + SS(X_1, X_2) = SS_{MLR Model} + SS_{Shared} \quad (54)$$

corresponds with regions 1, 2, and 4. The OVR model (Definition 3) now contains region 4, where the MLR model (Definition 2) did not. Now, $SS_{Total} - SS_{Error}$ does

equal SS_{Model} . The R^2 determined from the OVR model does actually include the shared variation and does equal Equation 45. Thus, whereas Equation 45 is inflated for the determination of R^2 when associated with the MLR model, it is now correct for the OVR model. Likewise, the F determined from Equation 46 is now appropriate for the OVR model by the result of Equation 54. As a consequence, the regression model and these summary statistics are now in agreement, which was not the case for the MLR model.

The ordered variable regression (OVR) can easily be performed within any statistical package using the following steps. [Although only presented for two predictors, the steps can easily be expanded to include any number of predictors. In addition, alternative orderings can easily be proposed, considered, and compared using the same method.] First, determine the order for considering the predictors. This is perhaps the hardest step, but most researchers have little or no trouble placing their predictors in some order based on logic, theory, convenience, and/or cost considerations. As a consequence, the research questions answered by the OVR model are arguably more consistent with real questions than those actually answered by the MLR model. For illustration, let X_1 be the predictor of primary interest. Second, obtain the residuals (X_{2res}) from the regression in which X_1 is the independent variable and X_2 is the dependent variable. The correlation between X_1 and these residuals will be zero. Thus the entire earlier section when the predictors are not related of this article applies. Third, perform the regression in which X_1 and X_{2res} are the predictors of the response variable Y . The result of this regression will be the OVR model expressed in Equation 49. Which will produce

$$R_{OVR Model}^2 = b_1^2 + r_{Y(X2.X1)}^2 = \text{Region 1} + \text{Region 4} + \text{Region 2} = R^2 \quad (55)$$

the R^2 value indicated in Equation 32 because $r_{Y(X2.X1)}^2$ is region 2 (from Equation 33) and

$$b_1^2 = r_{Y(X1.X2)}^2 + \frac{r_{12}^2}{1 - r_{12}^2} \left[r_{12} r_{Y(X1.X2)}^2 + r_{12} r_{Y(X2.X1)}^2 + 2 r_{Y(X1.X2)} r_{Y(X2.X1)} \right] \quad (56)$$

is region 1 + region 4 from Equations 33 and 37.

Numerical Example

To illustrate the points made in the two sections above for the SLR, MLR, and OVR models, and their corresponding results, a numerical example is presented in Table 2 for data in which $r_{12} = .469$. These results can be found by using the regression routine in any of the major statistical computer packages. The exception is that a general linear model routine needs to be run in order to obtain the sums of squares breakdown information specific to each predictor in the MLR and OVR contexts. Due to round off errors in the computation of X_{2res} , the actual correlation of X_1 and X_{2res} is $-5.2E-16$ instead of perfect zero.

At this point, it may seem that the OVR model is nothing more than hierarchical multiple regression analysis (HMR) or forward step regression using type I sum of squares. It is true that OVR and HMR share a common approach in that predictors are entered into the model sequentially and the additive contribution of each predictor can be reflected in the type I sum of squares. However, OVR differs from HMR in that the additive contribution of each predictor is reflected in both the type I sum of squares and the model coefficients. This is illustrated in Table 3. Of course, the OVR produces the same model as the HMR when predictors are not related.

It should be noted the concept of [what is referred to in this paper as] OVR was proposed by Woolf (1951) as a second method of calculating multiple linear regression. The novelty presented here is in the application of OVR as a method of regression modeling when faced with multicollinearity; guided by theory, OVR can be used to incrementally model the natural relatedness between predictors. As a consequence, OVR not only provides an alternative method of dealing with multicollinearity in a regression context, but more importantly, it allows the evaluation of research questions that assume or hypothesize hierarchical relatedness among predictors.

In summary, when the predictors are related, the coefficients of the SLR, MLR, HMR, and OVR models are not equal, but differ from one another in a predictable manner based on the amount of the relatedness between the two predictors. The data confirmed that the overall summary and test statistics (R^2 and F) associated with MLR are all inflated in relation to the model by the inclusion of the shared variance; as indicated in Equations 35, 45, and 46. In contrast, the data showed that these statistics are consistent with the OVR model which does included the shared variance. The implications and impacts of the results presented in Tables 2 and 3 will hold for any value of the relatedness between predictors that is different from zero.

BAIRD & BIEBER

Table 2. Comparison of Simple, Multiple, and Ordered Regression when the predictors are related

	SLR(X_1)	SLR(X_2)	MLR	OVR	Comments
Coefficient					
X_1	$b_1 = .505$		$c_1 = .389$	.505	$b_1 = r_{YX1}$, $b_1 \neq c_1$, OVR slope = b_1 , Equations 3, 27, 30, 49
X_2		$b_1 = .429$	$c_2 = .246$	.246	$b_2 = r_{YX2}$, $b_2 \neq c_2$, OVR slope = c_2 , Equations 4, 27, 30, 49
Sums of Squares					
X_1	18.854		8.753	18.854	Equations 23, 52
X_2		13.592	3.491	3.491	Equations 24, 52
Model	18.854	13.592	12.244	22.345	MLR (Equation 40), OVR (Equation 54)
Error	55.146	60.408	51.655	51.655	
Shared			10.101		Equations 41, 42
Total	74.000	74.000	74.000	74.000	MLR (Equation 39), OVR (Equation 50) For the SS to be additive, MLR must add in SS_{Shared}
R^2 Estimates					
X_1	.255		.118	.255	MLR (Equation 33), OVR (Equation 6)
X_2		.184	.047	.047	MLR, OVR (Equation 33)
R^2_{Model}			.137	.302	MLR (Equation 34), OVR (Equations 35, 55)
R^2_{Shared}			.165		MLR (Equations 37, 38), OVR is included in the model
R^2			.302		For the R^2 to be additive, MLR must add in R^2_{Shared}
F Statistics					
F_{Model}			8.533	15.573	MLR (Equation 47), OVR (Equation 46)
F_{Shared}			7.040		MLR (Equation 48)
F			15.573	15.573	Overall F value, MLR (Equation 46) must include F_{Shared}

*Note: $r_{12} = .469$, $r_{YX1} = .505$, $r_{YX2} = .429$, $r_{Y(X1.X2)} = .344$, $r_{Y(X2.X1)} = .217$, $n = 75$

THE GOLDBLOCKS DILEMMA

Table 3. Comparison of Hierarchical Multiple Regression, Forward Step Regression, and Ordered Variable Regression when predictors are related

	HMR	FSR	OVR	Comments
Coefficient				
X_1	.389	.389	.505	Note that the coefficients produced by HMR and FSR are identical to MLR from Table 3, but not OVR.
X_2	.246	.246	.246	
Sums of Squares (Type I)				
X_1	18.854	18.854	18.854	Note that the type I sum of squares matches across all models
X_2	3.491	3.491	3.491	
Model	22.345	22.345	22.345	
Error	51.655	51.655	51.655	
Shared				
Total	74.000	74.000	74.000	
R^2 Estimates				
X_1	.2548	.2548	.2548	
X_2	.0472	.0472	.0472	
R^2_{Model}	.302	.302	.302	

*Note: HMR and FSR models were run using SAS Software 9.3, using PROC REG, GLM and STEPWISE (SAS Institute Inc., Cary, NC)

Conclusion

Although there is no agreed upon definition of multicollinearity in the literature, the impacts of multicollinearity (or interrelatedness of the predictors) are straightforward, as presented in both of the implications and impacts sections of this article; regardless of the size of the relatedness. Specifically, when the predictors are interrelated, the model coefficients for the SLR models, the MLR model, and the OVR model are all different. What is more, the shared contribution resulting from the interrelatedness in MLR is included in the overall R^2 and F , but not in the model coefficients nor in the MLR model itself. However, this is not a problem for the OVR model as the same shared contribution is included in the R^2 and F as well as the model coefficients (and thus the OVR model).

Although rare, when no interrelatedness exists between the predictors, the SLR, MLR, and OVR coefficients and R^2 values are all consistent with each other. In addition, the MLR and OVR model coefficients, R^2 values, and F test statistic are all identical. In short, when interrelatedness does not exist between the predictors, all three definitions of joint contribution and their corresponding models are identical. This is summarized in Table 4.

Table 4. Summary of Simple, Multiple, and Ordered Regression when the predictors are related (regions of Figure 2)

	SLR Models		MLR Model	OVR Model	Comments		
	X_1	X_2					
Coefficients					Shared contribution is not included in MLR but is included in OVR		
X_1	1,4					1	1,4
X_2		2,4				2	2
R^2 Estimates					Shared contribution is included in R^2 and F for MLR, although MLR does not contain shared contribution. This is not a problem for the OVR		
R^2_{Model}	1,4	2,4				1,2	1,2,4
R^2	1,4	2,4				1,2,4	1,2,4
F Statistics							
F_{Model}	1,4	2,4				1,2	1,2,4
F	1,4	2,4				1,2,4	1,2,4

*Note. The shaded area indicates problems (the impacts) associated with the application of MLR.

Multicollinearity defined as the simple relatedness between predictors ($r_{12} \neq 0$) is a universal condition that exists within real data unless the predictors have been experimentally designed to be independent of each other. Consequently, the use of MLR will result in the impacts of multicollinearity as presented in this paper to an increasing degree as $|r_{12}|$ increases. Multicollinearity defined as a problematic condition that exists once $|r_{12}|$ increases beyond some threshold level, still results in the impacts presented in this paper. This second definition of multicollinearity is plagued by the need to ascertain a logical, reasonable, and appropriate threshold value. Although this is probably the more common of the two definitions, it presents the researcher with the hope of zero impact when in truth some degree of impact actually does exist (albeit smaller than the threshold amount). In either case, MLR results in a model that doesn't include the relatedness between the predictors.

OVR is presented as a method of modeling data when relatedness exists between predictors, a common issue in applied research. However, the behavior

THE GOLDBLOCKS DILEMMA

and generalizability of OVR with regard to other common applied issues, such as small sample size and departures of model assumptions, needs to be examined. Therefore, an essential next step in the research is to use Monte Carlo simulations to evaluate statistical power (of the corresponding F and t tests) and robustness of estimation and efficiency of OVR under conditions where asymptotic behavior often breaks down.

When faced with a regression problem with multiple related predictors, a researcher is confronted with the Goldilocks dilemma (see [Nestrick, 1962](#)). It is possible to address the problem from the perspective of the multiple application of simple regression (the papa bear solution which over includes the shared variance, [Equation 25](#)), from the perspective of multiple regression (the mama bear solution which doesn't include the shared variance, [Equation 34](#)) and from the perspective of order variable regression (the baby bear solution which appropriately considers the shared variance, [Equation 55](#)).

Acknowledgements

We would like to thank Jared Studyvin for his assistance throughout the development of this project and for his insightful review of the final manuscript.

References

- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69(3), 161-182. doi:10.1037/h0025471
- Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology*, 73(5), 592-616.
- Kerlinger, F. N. & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, Inc.
- Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). Homewood, IL: McGraw-Hill/Irwin.
- McNemar, Q. (1962). *Psychological statistics*. (3rd ed.). New York: Wiley.
- Nestrick, N. (Ed.). (1962). *The three bears and Goldilocks*. New York: Platt & Munk.
- Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). *Statistical package for the social sciences (SPSS)* (2nd ed.). New York: McGraw-Hill.
- Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw Hill.
- Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.
- Weisberg, S. (2005). *Applied Linear Regression* (3rd ed.). New York: Wiley.
- Wolf, B. (1951). Computation and interpretation of multiple regressions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(1), 100-119.

Estimation of Population Mean Using Exponential Type Imputation Technique for Missing Observations

Rajesh Singh
Banaras Hindu University
Varanasi, India

Hemant K. Verma
Banaras Hindu University
Varanasi, India

Prayas Sharma
University of Petroleum and
Energy Studies
Dehradun, India

Some imputation techniques are suggested for estimating the population mean when the data values are missing completely at random under a simple random sample without replacement scheme. Two classes of point estimators are proposed. The bias and mean squared error expressions of the proposed point estimators are derived up to first order of approximation. It has been shown that the proposed point estimators are more efficient than some existing point estimators due to Lee, Rancourt, and Sarndal (1994) and Singh and Horn (2000). Theoretical findings are supported by an empirical study based on five populations to show the superiority of the constructed estimators and methods of imputation over others.

Keywords: Missing data, imputation, bias, mean squared error, simple random sampling without replacement

---

Introduction

Missing data is a common and serious problem in survey sampling. Missing data naturally occurs in sample surveys when a few sampling units refuse to respond or are unable to participate in the survey. There are two types of non-responses which occur in surveys: unit non-response and item non-response. Unit non-response occurs when an eligible sample unit fails to participate in a survey because of failure to establish a contact or explicit refusal to cooperate. Item non-response occurs instead when a responding unit does not provide useful answers to particular items

---

Dr. Singh is an Assistant Professor in the Department of Statistics. Email him at: rsinghstat@gmail.com. Dr. Sharma, the corresponding author, is a Professor in the Department of Decision Sciences, College of Management and Economics Studies. Email at prayassharma02@gmail.com.

of the questionnaire. Such situations create missing data problem. The imputation is a well-defined methodology by virtue of which such problems can be unraveled.

In the literature several imputation techniques are available and discussed. Rubin (1976) addressed three concepts: observed at random (OAR), missing at random (MAR), and parametric distribution (PD). Rubin defined MAR as the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the value of the unobserved data. Heitjan and Basu (1996) distinguished the meaning of MAR and missing completely at random (MCAR) in a very nice way. The imputation technique is also applicable when information on auxiliary variable is available. Lee et al. (1994; 1995) used the information on an auxiliary variable for the purpose of imputation, Singh and Horn (2000) suggested a compromised method of imputation, Ahmed, Al-Titi, Al-Rawi, and Abu-Dayyeh (2006) suggested several new imputation based estimators that use the information on an auxiliary variable and compared their performances with the mean method of imputation, and Rao and Sitter (1995) used the imputation techniques for variance estimation under two phase sampling. Kadilar and Cingi (2008) and Diana and Perri (2010) also suggested some imputation techniques in case of missing data. In the present study we implicitly assume MCAR.

Let

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$$

be the population mean of study variable Y . A simple random sample without replacement (SRSWOR), s , of size n is drawn from $\Omega = \{1, 2, \dots, N\}$ to estimate the population mean $\bar{Y}$. Let r be the number of responding units out of sampled n , then the number of non-responding units is $(n - r)$. Let the set of responding units be denoted by R and that of non-responding units be denoted by R^c . For every unit $i \in R$, the value y_i is observed. However for the units $i \in R^c$, the y_i values are missing and imputed values are to be derived. We assume that imputation is carried out with the aid of a quantitative auxiliary variable x such that, the value of x for unit i is x_i , known and positive for every $i \in s$. In other words, the data $x_s = \{x_i : i \in s\}$ are known.

Some Available Methods of Imputation and Estimators

There are some classical methods of imputation which are commonly used and given as follows:

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

Mean Method of Imputation

In this method of imputation, the study variable y after imputation takes the form as

$$y_{.i} = \begin{cases} y_i, & i \in R \\ \bar{y}_r, & i \in R^c \end{cases} \quad (1)$$

and the point estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_s = \frac{1}{n} \sum_{i \in s} y_{.i} . \quad (2)$$

Thus, under this method of imputation, the point estimator of the population mean $\bar{Y}$ is

$$\bar{y}_m = \frac{1}{r} \sum_{i \in R} y_i = \bar{y}_r . \quad (3)$$

Lemma 1. The expression of Bias and Variance of the point estimator $\bar{y}_m$ is given as

$$\text{Bias}(\bar{y}_m) = 0 \quad (4)$$

$$V(\bar{y}_m) = \left(\frac{1}{r} - \frac{1}{N} \right) S_y^2 \quad (5)$$

where $S_y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})^2$.

Ratio Method of Imputation

Following the notations of Lee et al. (1994), in the case of single value imputation, if the i^{th} unit requires imputation, the value $\hat{bx}_i$ is imputed. Thus, the study variable y after imputation takes the form as

$$y_{\cdot i} = \begin{cases} y_i, & i \in R \\ \hat{b}x_i, & i \in R^c \end{cases} \quad (6)$$

where

$$\hat{b} = \frac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i}.$$

Under this method of imputation, the point estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_{\text{RAT}} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \quad (7)$$

where $\bar{x}_n = \frac{1}{n} \sum_{i \in S} x_i$, $\bar{x}_r = \frac{1}{r} \sum_{i \in R} x_i$, and $\bar{y}_r = \frac{1}{r} \sum_{i \in R} y_i$.

Lemma 2. The expression of Bias and Mean Square Error (MSE) of the point estimator $\bar{y}_{\text{RAT}}$ is given as

$$\text{Bias}(\bar{y}_{\text{RAT}}) = \bar{Y} \left(\frac{1}{r} - \frac{1}{n} \right) (C_x^2 - \rho C_y C_x), \quad (8)$$

$$\text{MSE}(\bar{y}_{\text{RAT}}) = \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n} \right) (S_y^2 + R_1^2 S_x^2 - 2R_1 S_{xy}), \quad (9)$$

where S_y^2 is defined as above and

$$S_x^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2, S_{xy} = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X})$$

$$R_1 = \frac{\bar{Y}}{\bar{X}}, C_y = \frac{S_y}{\bar{Y}}, C_x = \frac{S_x}{\bar{X}}, \rho = \frac{S_{xy}}{S_x S_y}$$

Compromised Method of Imputation

Singh and Horn (2000) proposed compromised imputation procedure. After imputation the study variable takes form as

$$y_{.i} = \begin{cases} \frac{\alpha n y_i}{r + (1-\alpha) \hat{b} x_i}, & i \in R \\ (1-\alpha) \hat{b} x_i, & i \in R^c \end{cases} \quad (10)$$

where α is a suitably chosen constant such that the variance of the resultant estimator is minimum. Here, we are also using information from imputed values for the responding units in addition to non-responding units.

Thus, under compromised method of imputation, the point estimator of the population mean $\bar{Y}$ is

$$\bar{y}_{\text{COMP}} = \alpha \bar{y}_r + (1-\alpha) \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \quad (11)$$

Lemma 3. The expression of Bias and MSE of the point estimator $\bar{y}_{\text{COMP}}$ is given as

$$\text{Bias}(\bar{y}_{\text{COMP}}) = \bar{Y} (1-\alpha) \left(\frac{1}{r} - \frac{1}{n} \right) (C_x^2 - \rho C_y C_x) \quad (12)$$

$$\begin{aligned} \text{MSE}(\bar{y}_{\text{COMP}}) = & \left(\frac{1}{n} - \frac{1}{N} \right) S_y^2 + \left(\frac{1}{r} - \frac{1}{n} \right) (S_y^2 + R_1^2 S_x^2 - 2R_1 S_{xy}) \\ & - \left(\frac{1}{r} - \frac{1}{n} \right) \bar{Y}^2 \alpha^2 C_x^2 \end{aligned} \quad (13)$$

where $\alpha_{\text{opt}} = 1 - \rho \frac{C_y}{C_x}$. Thus

$$\text{MSE}(\bar{y}_{\text{COMP}})_{\min} = \left[\left(\frac{1}{n} - \frac{1}{N} \right) - \left(\frac{1}{r} - \frac{1}{n} \right) \rho^2 \right] S_y^2 \quad (14)$$

Along similar lines, Ahmed et al. (2006) proposed several new imputation techniques by introducing some unknown parameters and hence proposed the corresponding estimators for estimating the finite population means $\bar{Y}$.

Proposed Imputation Methods and Corresponding Estimators

The following two imputation methods are suggested. After imputation for the first proposed imputation of technique, the study variable takes the form as

$$y_{.i} = \begin{cases} y_i, & i \in R \\ \frac{1}{n-r} \bar{y}_r \left[n \exp \left\{ \alpha \frac{\bar{X}^{\frac{1}{h}} - \bar{x}_r^{\frac{1}{h}}}{\bar{X}^{\frac{1}{h}} + (a-1)\bar{x}_r^{\frac{1}{h}}} \right\} - r \right], & i \in R^c \end{cases} \quad (15)$$

where a , h , and α are suitably chosen constants. We optimize α in such a way that the MSE of the resultant estimator is minimum. Thus we have the following theorem:

Theorem 1. Under the proposed method of imputation considered in (15), the point estimator of the population mean $\bar{Y}$ is given as

$$T_p = \bar{y}_r \exp \left\{ \alpha \frac{\bar{X}^{\frac{1}{h}} - \bar{x}_r^{\frac{1}{h}}}{\bar{X}^{\frac{1}{h}} + (a-1)\bar{x}_r^{\frac{1}{h}}} \right\}. \quad (16)$$

Proof:

$$T_p = \frac{1}{n} \sum_{i \in S} y_{.i} = \frac{1}{n} \left[\sum_{i \in R} y_{.i} + \sum_{i \in R^c} y_{.i} \right] \quad (17)$$

where R and R^c are the sets of responding and non-responding units in the sample s of size n .

Now putting the values from (15) into (17), the point estimator of population is obtained as mean $\bar{Y}$ as defined in (16), which completes the proof.

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

Table 1. Members of the class of estimators T_P

Estimators $\alpha = 1$		Constants a	h
$T_{P1} = \bar{y}_r \exp \left[\frac{\bar{X} - \bar{x}_r}{\bar{X}} \right]$	$T_{P5} = \bar{y}_r \exp \left[\frac{\bar{x}_r - \bar{X}}{\bar{X}} \right]$	1	1
$T_{P2} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}} - \sqrt{\bar{x}_r}}{\sqrt{\bar{X}}} \right]$	$T_{P6} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{x}_r} - \sqrt{\bar{X}}}{\sqrt{\bar{X}}} \right]$	1	2
$T_{P3} = \bar{y}_r \exp \left[\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r} \right]$	$T_{P7} = \bar{y}_r \exp \left[\frac{\bar{x}_r - \bar{X}}{\bar{X} + \bar{x}_r} \right]$	2	1
$T_{P4} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}} - \sqrt{\bar{x}_r}}{\sqrt{\bar{X}} + \sqrt{\bar{x}_r}} \right]$	$T_{P8} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{x}_r} - \sqrt{\bar{X}}}{\sqrt{\bar{X}} + \sqrt{\bar{x}_r}} \right]$	2	2

Because the point estimator proposed in (16) after imputing the missing values, belongs to a class of estimators. Some members of the proposed class of point estimator defined in (16) are shown in Table 1 for different choice of a , h , and α .

The study variable after imputation for the second proposed imputation of technique becomes

$$y_{.i} = \begin{cases} y_i, & i \in R \\ \frac{1}{n-r} \bar{y}_r \left[n \exp \left\{ \alpha \frac{\bar{X}^{\frac{1}{h}} - \bar{x}_n^{\frac{1}{h}}}{\bar{X}^{\frac{1}{h}} + (a-1)\bar{x}_n^{\frac{1}{h}}} \right\} - r \right], & i \in R^c \end{cases} \quad (18)$$

where a , h , and α are suitably chosen constants. We optimize α in such a way that the MSE of the resultant estimator is minimum. Thus we have the following theorem:

Theorem 2. Under the proposed method of imputation considered in (18), the point estimator of the population mean $\bar{Y}$ is given as

$$\bullet T_g = \bar{y}_r \exp \left\{ \alpha \frac{\bar{X}^{\frac{1}{h}} - \bar{X}_n^{\frac{1}{h}}}{\bar{X}^{\frac{1}{h}} + (a-1) \bar{X}_n^{\frac{1}{h}}} \right\} \quad (19)$$

Proof:

$$T_g = \frac{1}{n} \sum_{i \in S} y_{\cdot i} = \frac{1}{n} \left[\sum_{i \in R} y_{\cdot i} + \sum_{i \in R^c} y_{\cdot i} \right] \quad (20)$$

where R and R^c are the sets of responding and non-responding units in the sample, s , of size n .

Putting the values from (18) into (20), we get the form of the point estimator of population mean $\bar{Y}$ as defined in (19), which completes the proof.

Some members of the proposed class of point estimator defined in (19) are shown in Table 2 for different choices of a , h , and α .

Properties of the Estimators T_P and T_g

To obtain the bias and MSE expressions of the estimators to the first degree of approximation, we define

$$e_0 = \frac{\bar{y}_r - \bar{Y}}{\bar{Y}}, e_1 = \frac{\bar{x}_r - \bar{X}}{\bar{X}}, e_2 = \frac{\bar{x}_n - \bar{X}}{\bar{X}}$$

such that $E(e_i) = 0$; $i = 0, 1, 2$, and

$$\begin{aligned} E(e_0^2) &= \left(\frac{1}{r} - \frac{1}{N} \right) C_y^2, E(e_1^2) = \left(\frac{1}{r} - \frac{1}{N} \right) C_x^2, E(e_0 e_1) = \left(\frac{1}{r} - \frac{1}{N} \right) \rho C_y C_x, \\ E(e_2^2) &= \left(\frac{1}{n} - \frac{1}{N} \right) C_x^2, E(e_1 e_2) = \left(\frac{1}{n} - \frac{1}{N} \right) C_x^2, E(e_0 e_2) = \left(\frac{1}{n} - \frac{1}{N} \right) \rho C_y C_x \end{aligned}$$

Using above terminology, the bias and MSE of the proposed estimators are given below.

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

Table 2. Members of the class of estimators T_g

Estimators		Constants	
$\alpha = 1$	$\alpha = 1$	a	h
$T_{g1} = \bar{y}_r \exp \left[\frac{\bar{X} - \bar{X}_n}{\bar{X}} \right]$	$T_{g5} = \bar{y}_r \exp \left[\frac{\bar{X}_n - \bar{X}}{\bar{X}} \right]$	1	1
$T_{g2} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}} - \sqrt{\bar{X}_n}}{\sqrt{\bar{X}}} \right]$	$T_{g6} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}_n} - \sqrt{\bar{X}}}{\sqrt{\bar{X}}} \right]$	1	2
$T_{g3} = \bar{y}_r \exp \left[\frac{\bar{X} - \bar{X}_n}{\bar{X} + \bar{X}_n} \right]$	$T_{g7} = \bar{y}_r \exp \left[\frac{\bar{X}_n - \bar{X}}{\bar{X} + \bar{X}_n} \right]$	2	1
$T_{g4} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}} - \sqrt{\bar{X}_n}}{\sqrt{\bar{X}} + \sqrt{\bar{X}_n}} \right]$	$T_{g8} = \bar{y}_r \exp \left[\frac{\sqrt{\bar{X}_n} - \sqrt{\bar{X}}}{\sqrt{\bar{X}} + \sqrt{\bar{X}_n}} \right]$	2	2

Theorem 3. The Bias of the estimator T_P is given by

$$\text{Bias}(T_P) = \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y} \frac{\alpha}{ah} \left[\frac{1}{ah} \left(\frac{\alpha}{2} + a - 1 \right) C_x^2 - \rho C_y C_x \right] \quad (21)$$

and the MSE of the estimator T_P is given by

$$\text{MSE}(T_P) = \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 \left[C_y^2 + \frac{\alpha^2}{a^2 h^2} C_x^2 - 2 \frac{\alpha}{ah} \rho C_y C_x \right], \quad (22)$$

where the optimum value of α is given by

$$\alpha_{\text{opt}} = ah\rho \frac{C_y}{C_x}$$

Proof: Expressing the estimator T_P in terms of the e 's, we have

$$T_P = \bar{Y} (1 + e_0) \exp \left[\alpha \left\{ \frac{1 - (1 + e_1)^{\frac{1}{h}}}{1 + (a-1)(1 + e_1)^{\frac{1}{h}}} \right\} \right]$$

$$\begin{aligned}
&\approx \bar{Y}(1+e_0) \exp \left[\frac{-\alpha e_1}{ah} \left(1 + \frac{e_1}{h} - \frac{e_1}{ah} \right)^{-1} \right] \\
&\approx \bar{Y}(1+e_0) \exp \left(\frac{-\alpha e_1}{ah} \right) \exp \left(\frac{\alpha(a-1)}{a^2 h^2} e_1^2 \right) \\
&\approx \bar{Y}(1+e_0) \left[1 - \frac{\alpha e_1}{ah} + \frac{\alpha(a-1)}{a^2 h^2} e_1^2 + \frac{1}{2} \frac{\alpha}{a^2 h^2} e_1^2 \right] \\
(T_P - \bar{Y}) &\approx \left[e_0 - \frac{\alpha}{ah} e_1 - \frac{\alpha(a-1)}{a^2 h^2} e_1^2 + \frac{1}{2} \frac{\alpha}{a^2 h^2} e_1^2 \right]
\end{aligned}$$

Taking expectation on both sides, we get the bias expression of estimator T_P as

$$\text{Bias}(T_P) = \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y} \frac{\alpha}{ah} \left[\frac{1}{ah} \left(\frac{\alpha}{2} + a - 1 \right) C_x^2 - \rho C_y C_x \right]$$

To find the MSE of the estimator T_P , we have

$$\begin{aligned}
\text{MSE}(T_P) &= E(T_P - \bar{Y})^2 \\
&\approx \bar{Y}^2 E \left(e_0 - \frac{\alpha}{ah} e_1 \right)^2 \\
&\approx \bar{Y}^2 \left(E(e_0^2) + \frac{\alpha^2}{a^2 h^2} E(e_1^2) - \frac{\alpha}{ah} E(e_0 e_1) \right) \\
\text{MSE}(T_P) &= \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 \left[C_y^2 + \frac{\alpha^2}{a^2 h^2} C_x^2 - 2 \frac{\alpha}{ah} \rho C_y C_x \right]
\end{aligned}$$

Partially differentiating above equation with respect to α and equating to zero, we have

$$\frac{\partial(\text{MSE}(T_P))}{\partial \alpha} = \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 \left[2 \frac{\alpha}{a^2 h^2} C_x^2 - 2 \frac{1}{ah} \rho C_y C_x \right] = 0$$

Simplifying the above equation, we get the optimum value of α as

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

$$\alpha_{\text{opt}} = ah\rho \frac{C_y}{C_x}$$

Theorem 4. The Bias of the estimator T_g is given by

$$\text{Bias}(T_g) = \left(\frac{1}{n} - \frac{1}{N} \right) \bar{Y} \frac{\alpha}{ah} \left[\frac{1}{ah} \left(\frac{\alpha}{2} + a - 1 \right) C_x^2 - \rho C_y C_x \right] \quad (23)$$

and the MSE of the estimator T_g is given by

$$\text{MSE}(T_g) = \bar{Y}^2 \left[\left(\frac{1}{r} - \frac{1}{N} \right) C_y^2 + \left(\frac{1}{n} - \frac{1}{N} \right) \left(\frac{\alpha^2}{a^2 h^2} C_x^2 - 2 \frac{\alpha}{ah} \rho C_y C_x \right) \right], \quad (24)$$

where the optimum value of α is given by

$$\alpha_{\text{opt}} = ah\rho \frac{C_y}{C_x}$$

Proof: The above theorem can be proved in a similar way to the proof of Theorem 3.

Efficiency Comparison

Estimator T_P is more efficient than estimator $\bar{y}_m$ if

$$V(\bar{y}_m) - \text{MSE}[T_P(\min)] > 0.$$

But

$$\begin{aligned} \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 C_y^2 - \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 (1 - \rho^2) C_y^2 &> 0 \\ \left(\frac{1}{r} - \frac{1}{N} \right) \bar{Y}^2 \rho^2 C_y^2 &> 0 \end{aligned}$$

since $\frac{1}{r} > \frac{1}{N}$. Therefore, T_P is more efficient than $\bar{y}_m$. Similarly,

$$\text{MSE}(\bar{y}_{\text{RAT}}) - \text{MSE}[T_P(\min)] > 0 \text{ if } \left(\frac{1}{r} - \frac{1}{n}\right)(C_x^2 - 2\rho C_y C_x) + \left(\frac{1}{r} - \frac{1}{N}\right)\rho^2 C_y^2 > 0$$

$$\text{MSE}(\bar{y}_{\text{COMP}}) - \text{MSE}[T_P(\min)] > 0 \text{ if } \left(\frac{1}{r} - \frac{1}{n}\right) < \left(\frac{1}{r} - \frac{1}{N}\right)\rho^2$$

$$\text{MSE}[T_g(\min)] - \text{MSE}[T_P(\min)] > 0 \text{ if } \frac{1}{r} > \frac{1}{n}$$

Thus, from the above results, we can say that the estimator T_P is more efficient than other estimators.

Empirical Study

Five populations, A, B, C, D, and E, are considered. Population A is the artificial population of size $N=200$ from Shukla, Thakur, Pathak, and Rajput (2009), population B is from Ahmed et al. (2006), population C is from Dass (1988), population D is from Murthy (1967, p. 228), and population E is from Singh, Singh, and Kumar (1976, p. 126) with parameters as given in Table 3.

Let $n = 40$, $r = 35$ for population A, $n = 200$, $r = 180$ for population B, $n = 80$, $r = 72$ for population C, $n = 23$, $r = 20$ for population D, and $n = 6$, $r = 5$ for population E respectively. Then the bias and MSE of the proposed point estimators are given in Table 4 and Table 5 for populations A, B, C, D and E respectively.

Table 3. Parameters for study populations

Population	Parameters							
	N	$\bar{Y}$	$\bar{X}$	S_y^2	S_x^2	ρ	C_y	C_x
A	200	42.485	18.515	199.0598	48.5375	0.865200	0.37630	0.33210
B	8306	253.750	343.316	338006.0000	862017.0000	0.522231	2.70436	2.29116
C	278	39.070	25.110	3199.2400	1660.0200	0.720000	1.44770	1.62260
D	80	5182.640	285.130	3370161.0000	73129.9400	0.920000	0.35420	0.94840
E	17	33.290	40.060	287.8600	458.3500	0.720000	0.50970	0.54990

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

Table 4. Biases of estimators

Estimators	Populations				
	A	B	C	D	E
$\bar{y}_m$	0.0000	0.0000	0.0000	0.0000	0.0000
$\bar{y}_{\text{RAT}}$	0.0051	0.5749	0.0511	19.9568	0.1117
$\bar{y}_{\text{COMP}}$	0.0039	0.2543	0.0328	6.8572	0.0745
$T_P(\text{min})$	-0.0413	-0.9872	-0.8184	-10.3199	-0.3164
$T_g(\text{min})$	-0.0351	-0.8863	-0.1890	-8.5252	-0.2417

Table 5. MSEs of estimators

Estimators	Populations				
	A	B	C	D	E
$\bar{y}_m$	4.6921	1837.1169	32.9258	126381.0375	40.6391
$\bar{y}_{\text{RAT}}$	4.2110	1867.2341	31.3361	175668.0261	36.9018
$\bar{y}_{\text{COMP}}$	4.1599	1785.9043	30.6224	107777.7488	35.6648
$T_P(\text{min})$	1.1798	1336.0843	15.8571	19412.1274	19.5718
$T_g(\text{min})$	2.8938	1387.2968	18.9807	38015.4161	24.5460

Tables 4 and 5 exhibits the bias and MSE of different point estimators and it has been observed from the tables that the estimators based on auxiliary information are more efficient than the one which does not use the auxiliary information such as $\bar{y}_m$ to overcome the imputation problems. Both the proposed classes of estimators T_P and T_g are more efficient than the estimators, $\bar{y}_m$, $\bar{y}_{\text{RAT}}$ and $\bar{y}_{\text{COMP}}$, scrupulously, T_P has minimum MSE among all the estimators considered here.

Conclusion

Two imputation techniques are suggested using auxiliary information followed by two class of estimators for estimating the population mean in case of data values are MCAR under a SRSWR scheme. In addition, some new members are also generated from two proposed class of estimators using the suitable values of constants. The minimum biases and mean square errors of the proposed class of estimators were determined up to the first order of approximation. It was established theoretically and empirically that the proposed class of estimator performs best among the other estimators considered, and consequently the

corresponding (first proposed) method of imputation is better than the other existing methods and may be recommended for further use.

References

- Ahmed, M. S., Al-Titi, O., Al-Rawi, Z., & Abu-Dayyeh, W. (2006). Estimation of a population mean using different imputation methods. *Statistics in Transition*, 7(6), 1247-1264.
- Dass, A. K. (1988). *Contributions to the theory of sampling strategies based on auxiliary information* (Unpublished doctoral dissertation). Bidhan Chandra Agricultural University, West Bengal, India.
- Diana, G., & Perri, P. F. (2010). Improved estimators of the population mean for missing data. *Communications in Statistics – Theory and Methods*, 39(18), 3245-3251. doi: 10.1080/03610920903009400
- Heitjan, D. F., & Basu, S. (1996). Distinguishing “missing at random” and “missing completely at random”. *The American Statistician*, 50(3), 207-213. doi: 10.1080/00031305.1996.10474381
- Kadilar, C., & Cingi, H. (2008). Estimators for the population mean in the case of missing data. *Communications in Statistics – Theory and Methods*, 37(14), 2226-2236. doi: 10.1080/03610920701855020
- Lee, H., Rancourt, E., & Sarndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, 10(3), 231-243. Retrieved from: <http://www.jos.nu/Articles/abstract.asp?article=103231>
- Lee, H., Rancourt, E., & Sarndal, C. E. (1995). Variance estimation in the presence of imputed data for the generalized estimation system. In *Proceedings of the Survey Research Methods Section*, American Statistical Association, 384-389. Retrieved from: <http://www.amstat.org/sections/srms/Proceedings/>
- Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta, India: Statistical Publishing Society.
- Rao, J. N. K., & Sitter, R. R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82(2), p. 453-460. doi: 10.1093/biomet/82.2.453
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-593. doi: 10.1093/biomet/63.3.581

EXPONENTIAL TYPE IMPUTATION TECHNIQUE

Shukla, D., Thakur, N. S., Pathak, S., & Rajput, D. S. (2009). Estimation of mean under imputation of missing data using factor-type estimator in two-phase sampling, *Statistics in Transition*, 10(3), 397-414.

Singh, D., Singh, P., & Kumar, P. (1976). Handbook on Sampling Methods. New Delhi, India: Indian Agricultural Statistics Research Institute (ICAR).

Singh, S., & Horn, S. (2000). Compromised imputation in survey sampling. *Metrika*, 51(3), 266-276. doi: [10.1007/s001840000054](https://doi.org/10.1007/s001840000054)

Compound Identification Using Penalized Linear Regression on Metabolomics

Ruiqi Liu

University of Louisville
Louisville, KY

Dongfeng Wu

University of Louisville
Louisville, KY

Xiang Zhang

University of Louisville
Louisville, KY

Seongho Kim

Wayne State University
Detroit, MI

Compound identification is often achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity. Because the number of compounds in the reference library is much larger than the range of mass-to-charge ratio (m/z) values, so that the data become high dimensional data suffering from singularity. For this reason, penalized linear regressions such as ridge regression and the lasso are used instead of the ordinary least squares regression. Furthermore, two-step approaches using the dot product and Pearson's correlation along with the penalized linear regression are proposed in this study.

Keywords: Compound identification, mass spectral similarity, metabolomics, penalized linear regression

---

Introduction

One of the critical analyses on GC-MS data is compound identification, and it is often achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity (Stein & Scott, 1994). To improve the accuracy of compound identification, various algorithms measuring mass spectral similarity scores have been developed, such as dot product (Tabb, MacCoss, Wu, Anderson, & Yates, 2003; Beer, Barnea, Ziv, & Admon, 2004; Craig, Cortens, Fenyo, & Beavis, 2006; Frewen, Merrihew, Wu, Noble, &

Ruiqi Liu is a PhD student in the Department of Bioinformatics and Biostatistics. Email at: r0liu009@louisville.edu. Dr. Wu is an Associate Professor in the Department of Bioinformatics and Biostatistics. Email at: dongfeng.wu@louisville.edu. Dr. Zhang is a Professor in the Department of Chemistry. Email at: xiang.zhang@louisville.edu. Dr. Kim is the corresponding author and an Assistant Professor in the Department of Oncology. Email at: kimse@karmanos.org.

MacCoss, 2006), composite similarity (Stein & Scott, 1994), probability-based matching system (Atwater, Stauffer, McLafferty, & Peterson, 1985), Hertz similarity index (Hertz, Hites, & Biemann, 1971), normalized Euclidean distance (L2-norm) (Rasmussen & Isenhour, 1979; Stein & Scott 1994; Julian, Higgs, Gygi, & Hilton, 1998), absolute value distance (L1-norm) (Rasmussen & Isenhour, 1979; Beer et al., 2004), Fourier and wavelet-based composite similarity (Koo, Zhang, & Kim, 2011), and mixture partial and semi-partial correlation measures (Kim et al., 2012).

Because some compounds have mass spectral information that is similar to that of other compounds, an experimental query spectrum of these compounds is often matched to multiple mass spectra in the reference library with high similarity scores, impeding the high confidence compound identification. That is, the mass spectral similarity score of a true positive pair does not always have the top ranked score, and it is instead ranked as the second- or the third-highest similarity score with an ignorable difference from the top-ranked score.

In order to avoid the aforementioned issue, Kim et al. (2012) developed a novel similarity measure using partial and semi-partial correlations. The partial correlation can be seen as the pure relationship between two random variables after adjusting the effect of other random variables. On the other hand, the semi-partial correlation eliminates the effect of a fraction of other random variables, just adjusting the effect of one random variable from a total of two random variables. When it comes to compound identification, these partial and semi-partial correlations can be applied to calculate the mass spectral similarity score. By removing the effect of other mass spectra over the two mass spectra of interest, the unique relationship between the mass spectra can be extracted. Using partial and semi-partial correlations can obtain high accuracy of compound identification. Indeed, Koo, Kim, and Zhang (2013) recently compared among existing spectral similarity measures in terms of compound identification and concluded that mixture semi-partial correlation measure outperforms others. However, the performance of this method suffers from expensive calculation because the data are ultra-high-dimensional, which propels us to search for an alternative for compound identification.

Another way for compound identification is to use the multiple ordinary linear regression-based methods. In the context of linear regression, the response variable is an experimental mass spectrum (i.e., query) and all the compounds in the reference library are the independent variables. Each regression coefficient reflects the strength of their relationships with the response variable, so we could match the experimental compound with the reference compound which shows the strongest

connection. In particular, the coefficients of the multiple ordinary linear regressions are proportional to the semi-partial correlation coefficient, meaning that both methods will give us the same result if the maximal coefficient is considered only. In other words, the ordinary linear regression is a great alternative to the semi-partial correlation-based compound identification.

However, it is not feasible to apply ordinary linear regression in compound identification for two reasons. First, our data are high-dimensional data. The size of a reference library is much larger than the range of mass-to-charge ratio (m/z) values, and the number of variables becomes much larger than the number of samples so that the ordinary linear regression will suffer from singularity. Second, it is possible that different compounds have identical mass spectra, such as isomers. Note that isomers are compounds with the same molecular formula but different chemical structures. Because of the existence of isomers, several predictors are highly correlated to each other so that their correlation coefficients become almost one. This also causes ordinary linear regression to suffer from singularity.

In order to elude this difficulty, a penalized linear regression is introduced for the compound identification. Penalized linear regression can deal with high-dimensional data, and it is a trade-off between unbiasedness and a smaller estimation variance by putting a penalty constraint on coefficients. Different types of constraints will result in the lasso and ridge regression, which have L1-norm and L2-norm penalties, respectively. To improve the performance of penalized linear regression, two-step approaches are introduced using widely used mass spectral similarity scoring methods, either dot product or Pearson's correlations as the first step, and then penalized linear regression as the second step. Using the NIST mass spectral library, the performance of the proposed penalized linear regression approaches and two-step approaches with the dot product and Pearson's correlation are compared in terms of the accuracy of compound identification.

Methodology

Mass spectrum matching-based compound identification is achieved by matching the experimental mass spectra to the mass spectra stored in a reference library based on mass spectral similarity. In other words, all pairwise similarity scores between an experimental mass spectrum and each of the library mass spectra are first calculated. The compound whose library mass spectrum has the highest mass spectral similarity score is considered as the most probable compound that generated the experimental mass spectrum. Each mass spectrum is composed of

m/z values and their intensities. The intensities are used for calculation of the spectral similarity scores.

In this study, the spectral similarity between experimental mass spectrum and each of the reference spectra is calculated. A reference compound is considered as the compound given rise to the experimental spectrum if its reference spectrum has the best similarity with the experimental spectrum. The following methods are applied to calculate the similarity scores between the experimental mass spectrum and each of the reference spectra:

Dot Product

The dot product, which is also known as the cosine correlation (Stein & Scott, 1994), was used to obtain the cosine of the angle between two sequences of intensities, $\mathbf{x} = (x_i)_{i=1,\dots,n}$ and $\mathbf{y} = (y_i)_{i=1,\dots,n}$. It is defined as

$$S = S(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}, \quad (1)$$

where $\mathbf{x}^T \mathbf{y} = \sum_{i=1}^n x_i y_i$ and $\|\mathbf{x}\| = \left(\sum_{i=1}^n x_i^2 \right)^{1/2}$. We calculate the dot product of mass spectra for each experimental compound and each reference compound, and a greater value of S in (1) indicates a higher chance that the reference compound is the compound that generated the experimental mass spectrum.

Ridge Regression

Ridge regression is a shrinkage method which imposes a penalty on the size of regression coefficients. The ridge coefficients minimize a penalized residual sum of squares,

$$\beta^{\text{ridge}} = \arg \min_{\beta} \left\{ \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}, \quad (2)$$

where p is the number of variables (e.g., compounds or metabolites), N is the number of observations (e.g., intensities or m/z values), and $\lambda \geq 0$, which is a complexity parameter and controls the amount of shrinkage. A larger value of λ results in a great amount of shrinkage. The coefficients are shrunk toward zero (and

each other) (Hastie, Tibshirani, & Friedman, 2009). A well-known equivalent method is to solve the following problem, which makes the size constraint on the parameters explicit:

$$\beta^{\text{ridge}} = \arg \min_{\beta} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2, \quad (3)$$

subject to $\sum_{j=1}^p \beta_j^2 \leq t$. Note that there is a one-to-one correspondence between the parameters λ and t .

For ridge regression, we can also write the above criterion in matrix form, the ridge regression can be easily solved as

$$\beta^{\text{ridge}} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (4)$$

where $\mathbf{I}$ is the $p \times p$ identity matrix. In our case, $p \gg N$, so use the singular-value decomposition of $\mathbf{X}$, $\mathbf{X} = \mathbf{U} \mathbf{D} \mathbf{V}^T = \mathbf{R} \mathbf{V}^T$ to calculate the coefficients, where $\mathbf{V}$ is $p \times N$ with orthonormal columns, $\mathbf{U}$ is $N \times N$ orthogonal, and $\mathbf{D}$ is a diagonal matrix with elements $d_1 \geq d_2 \geq \dots \geq d_N \geq 0$. The matrix $\mathbf{R}$ is $N \times N$ with rows r_i^T . Replacing $\mathbf{X}$ by $\mathbf{R} \mathbf{V}^T$, we have

$$\beta^{\text{ridge}} = \mathbf{V} (\mathbf{R}^T \mathbf{R} + \lambda \mathbf{I})^{-1} \mathbf{R}^T \mathbf{y}. \quad (5)$$

The Lasso

The lasso (Least Absolute Shrinkage and Selection Operator), which was first proposed by Tibshirani (1996), is a shrinkage method like ridge, but it has subtle and important differences from the ridge regression. The lasso is a penalized least squares procedure that minimizes residual sum of squares (RSS) subject to the non-differentiable constraint expressed in terms of the L1 norm of the coefficients (Kyung, Gill, Ghosh, & Casella, 2010). That is, the lasso estimator is given by

$$\beta^{\text{lasso}} = \arg \min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^N \left(y_i - \beta_0 - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}. \quad (6)$$

This L1 norm constraint makes the solutions nonlinear in the y_i , resulting in no analytical solution different from ridge regression.

Two-Step Approach

To maximize the performance of compound identification and also reduce the data dimensionality, the two-step approaches are proposed by combining the dot product, Pearson's correlation, and penalized linear regression. In this procedure, the first step is made to precede the first match. Then, select a certain amount of the best matches based on the result of the first step and use them to conduct the second step which is penalized linear regression.

Dot product and lasso/ridge regression

In this two-step approach, after calculating the dot product of mass spectra for all experimental mass spectra and reference mass spectra, rank the results of dot product and choose N reference compounds with top N largest dot product values. Then conduct the lasso or ridge regression with only these N reference compounds. The flowchart is shown in [Figure 1](#).

Pearson's correlation and lasso/ridge regression

In this case, after calculating the Pearson's correlation coefficients of an experimental spectrum and all reference spectra, sort the correlation coefficients in descending order and calculate their $(1 - \alpha)\%$ confidence intervals. Then, check if there is overlap between two adjacent intervals from the top compounds and stop at the N^{th} compound, if there is no overlap between the N^{th} interval and $(N + 1)^{\text{th}}$ interval. By doing so, select N reference compounds and then conduct the lasso/ridge regression only with these N reference compounds. [Figure 2](#) shows the flow chart.

Figure 1. Workflows of the proposed two-step approach using dot product

Figure 2. Workflows of the proposed two-step approach using Pearson's correlation along with the lasso/ridge regression

Data

The National Institute of Standards and Technology (NIST) Chemistry WebBook service provides users with chemical and physical information for chemical compounds, including mass spectra generated by electron ionization mass spectrometry (Linstrom & Mallard, 2001). The mass spectra recorded in the NIST main mass spectrometry database and repetitive database were used as the reference mass spectra and experimental mass spectra, respectively. For our reference library, the mass spectra of 2739 compounds were extracted from NIST Chemistry WebBook database. The fragment ion m/z values ranged from 1 to 1036 with a bin size of 1. The experimental library contains 1530 mass spectra of compounds

extracted from the repetitive database. Because it was assumed the NIST library has the mass spectrum information for all the experimental compounds, all the compounds that were not present in the NIST main library were removed from the repetitive library.

Performance Evaluation

Each compound in the NIST database was assigned to a unique Chemical Abstract Service (CAS) registry number. To evaluate the performance of compound identification of each similarity measure, calculate the identification accuracy. The accuracy is the proportion of the spectra identified correctly in query data. In other words, if a pair of unknown and reference spectra have the same CAS index, we consider this pair as the correct match and if otherwise as the incorrect match. Then by counting all the correct matches, the accuracy of identification can be calculated by

$$\text{accuracy} = \frac{\text{number of spectra matched correctly}}{\text{number of spectra queried}} \quad (7)$$

Software

All the statistical analyses are performed using statistical software R version 2.15.3. The comparison of ridge regression and the lasso is performed by the R package *glmnet*.

Results

The penalized regressions, lasso and ridge regression, were conducted using R package *glmnet* to compare the identification results. In order to find a proper range of the shrinkage factor λ , the shrinkage factor was initially varied widely from 0.0001 to 1000000 and accuracy was calculated for each method. Figure 3(a) shows accuracy along with different shrinkage factor values for these two penalized linear regressions. The accuracy trend for the lasso is very different from that of ridge regression. For larger values of λ , accuracy tends to be a constant for each regression. However, accuracy for the lasso tends to be zero, while the ridge regression levels off at 89.20%. Based on this analysis, the shrinkage factors ranged from 0.10 to 5000 were focused on and then applied the lasso and ridge regression, respectively, to further check the specific trends of each regression.

Figure 3. Accuracy vs. shrinkage factor λ . Plot (a) is for the lasso and ridge regression using the wide range of λ . Plots (b) and (c) are for the ridge regression and lasso, respectively, using the smaller range of λ .

The Lasso

After conducting the lasso regression between query data and reference data with 100 different shrinkage factors λ (range from 0.10 to 5000), correct matches and accuracy were calculated. Figure 3(c) displays the change of accuracy corresponding to different shrinkage factor values. After a further check, the best accuracy for the lasso is 91.50% when $\lambda = 4646.47$. This accuracy is higher than the highest accuracy from ridge regression.

Two-Step Approach

Dot product and the lasso/ridge regression

The two-step approach, dot product and the lasso/ridge regression were performed to optimize the performance of compound identification, and to find the relationship between accuracy and different rank levels as well as λ values. A total of 12 different rank levels ranging from 25 to 300 were chosen. For λ , 100 values ranging from 0.10 to 5000 were used, which is the same with the identification using the lasso and ridge regression. Table 1 lists the analysis results. The results for this two-step approach are not so clear to interpret, so a contour plot (Figure 4) is used to show the relationship among accuracy, rank levels, and shrinkage factors for both the lasso and ridge regression.

In Figure 4, the green color indicates relatively low accuracy, while white and pink indicate relatively high accuracy. The highest accuracy, 90.20%, appears at rank level = 25 and $\lambda = 0.10$, which is shown as a red point in the left plot of Figure 4. The other four red points in the left plot of Figure 4 also have relatively high accuracy. Comparing with ridge regression only, we can see that this two-step approach performs better than the ridge regression only (accuracy = 90.20% vs. 89.74%). In general, we can also see the following trend: when the shrinkage factor (λ) increases, the corresponding rank needs to be increased in order to achieve better identification accuracy.

Table 1. Top 5 best accuracies and corresponding shrinkage factors for the dot product and the lasso/ridge regression

Method	Rank	Shrinkage factor (λ)	Number of query	Number of correct matches	Accuracy
Dot Product and Ridge	25	0.10	1530	1380	90.20%
	100	202.12	1530	1380	90.20%
	100	303.12	1530	1380	90.20%
	250	505.14	1530	1380	90.20%
	275	555.64	1530	1380	90.20%
Dot Product and Lasso	200	3838.41	1530	1395	91.18%
	300	1363.71	1530	1395	91.18%
	300	1414.21	1530	1395	91.18%
	300	1464.72	1530	1395	91.18%
	300	1515.22	1530	1395	91.18%

Figure 4. Accuracy of two-step approach using dot product and ridge (left) and the lasso regression (right)

The right plot of Figure 4 displays the relationship among accuracy, rank levels, and λ values for the two-step approach using the dot product and the lasso regression. The highest accuracy 91.18% appears at rank level = 200 and $\lambda = 3838.41$, which are shown as a red point in the plot. Comparing to the identification using the lasso only, this two-step approach has no improvement in accuracy, which is different from the two-step approach using ridge regression.

Pearson's correlation and the lasso/ridge regression

For the Pearson's correlation and penalized linear regression two-step approach, we intend to find the relationship among accuracy, different confidence levels, and λ values. The α levels of 0.01, 0.025, 0.05, and 0.1 were chosen, along with 100 shrinkage factor (λ) values ranging from 0.10 to 5000. The top 5 highest accuracies and corresponding shrinkage factors are shown in Table 2.

The best accuracies for this two-step approach using the lasso and ridge all appear at $\alpha = 0.1$, which are 89.41% (ridge regression) and 77.91% (the lasso). However, in this two-step approach, the lasso regression does not seem as good as the ridge regression. The contour plots are shown in Figure 5.

The relationship of accuracy, α levels, and λ values in this two-step approach seems much clearer. In the left plot of Figure 5, when the shrinkage factor (λ) is greater than a certain value (around 300), it does not influence the accuracy so much.

PENALIZED-BASED COMPOUND IDENTIFICATION

The red points, which indicate the best accuracies, all appear at $\alpha=0.1$, making a red vertical line.

Table 2. Top 5 best accuracies and corresponding shrinkage factors for Pearson's correlation and the lasso/ridge regression

Method	α	Shrinkage factor (λ)	Number of query	Number of correct matches	Accuracy
Dot Product and Ridge	0.1	101.11	1530	1368	89.41%
	0.1	353.63	1530	1368	89.41%
	0.1	404.13	1530	1368	89.41%
	0.1	454.64	1530	1368	89.41%
	0.1	505.14	1530	1368	89.41%
Dot Product and Lasso	0.1	0.10	1530	1192	77.91%
	0.1	50.60	1530	1192	77.91%
	0.1	101.11	1530	1192	77.91%
	0.1	151.61	1530	1192	77.91%
	0.1	202.12	1530	1192	77.91%

Figure 5. Accuracy of two-step approach using Pearson's correlation and ridge (left) and the lasso regression (right)

The relationship among accuracy, α levels, and λ values in Pearson's correlation and the lasso two-step approach is similar to that when ridge regression is used, as can be seen in the right plot of Figure 5. As in the two-step approach

using ridge regression, the red points all appear at $\alpha=0.1$, which make a red vertical line. The selection of λ value does not influence the accuracy, although it is clear that a greater α level results in higher accuracy.

Table 3. Compound identification methods and their performance.

Method	Lambda	Rank (Alpha)	Accuracy (%)
Dot Product	--	--	89.54
Pearson's Correlation	--	--	89.54
Ridge	1363.71	--	89.74
Lasso	4646.47	--	91.50
Dot Product and Ridge	0.10	25.0	90.20
Pearson's Correlation and Ridge	353.63~858.67	0.1	89.41
Dot Product and Lasso	3838.41	200.0	91.18
	1363.71~1515.22	300.0	
Pearson's Correlation and Lasso	0.10~960.00	0.1	77.91

The Best Performance

The performance of four compound identification methods involving penalized linear regression were tested. In addition, previously widely used methods were included. Table 3 shows these new methods and their best performance (accuracy), including the corresponding shrinkage factor (λ) value, rank selection (for dot product and the lasso/ridge regression two-step approach), and alpha selection (for Pearson's correlation and the lasso/ridge regression two-step approach). The performance of the dot product and Pearson's correlation in compound identification are also listed. Overall, the lasso only performs the best among other approaches (accuracy = 91.50%, line 4 in Table 3).

Conclusion

New approaches for compound identification were proposed using penalized linear regressions, and further two-step approaches are introduced. In particular, an alternative to the semi-partial correlation-based approach using multiple linear regressions was pursued.

From the results using a small data set, it can be seen that the lasso achieves the highest accuracy of compound identification, which is 91.50% with λ of 4646.5,

resulting in 1% greater accuracy than that of the dot product. Nevertheless, the accuracy for the lasso is highly related to the selection of shrinkage factor λ , so we have to tune up the shrinkage factor, such as using cross-validation, when using the lasso for compound identification. This additional work will result in a longer calculation time. Although ridge regression shows a worse accuracy than the lasso, its property that accuracy becomes constant after a certain λ value makes the ridge regression a better choice in terms of computational expense. In addition, the two-step approach using the dot product and the lasso has accuracy 91.18 %, which is similar to that of the lasso only. Because the dot product reduces the size of library, the following lasso regression becomes much inexpensive than the lasso regression only in terms of computational time. In this regard, this method could be a best alternative to the lasso regression only to achieve a higher accuracy.

Furthermore, the same data used here were applied to the mixture semi-partial correlation approach with the mixture weight of 0.7 and the rank of 100 ([Kim et al. 2012](#)), resulting in a slightly better performance than that of the lasso only with 92.9% of identification accuracy. Although the two-step approach using Pearson's correlation and the lasso/ridge regression has no improvement in identification accuracy, it shows that the shrinkage factor selection has no effect upon the accuracy of compound identification, which means that there should be no concern about the selection of shrinkage factors.

References

- Atwater, B. L., Stauffer, D. B., McLafferty, F. W., & Peterson, D. W. (1985). Reliability ranking and scaling improvements to the probability based matching system for unknown mass spectra. *Analytical Chemistry*, 57(4), 899-903. doi: [10.1021/ac00281a028](https://doi.org/10.1021/ac00281a028)
- Beer, I., Barnea, E., Ziv, T., & Admon, A. (2004). Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, 4(4), 950-960. doi: [10.1002/pmic.200300652](https://doi.org/10.1002/pmic.200300652)
- Craig, R., Cortens, J. C., Fenyo, D., & Beavis, R. C. (2006). Using annotated peptide mass spectrum libraries for protein identification. *Journal of Proteome Research*, 5(8), 1843-1849. doi: [10.1021/pr0602085](https://doi.org/10.1021/pr0602085)
- Frewen, B. E., Merrihew, G. E., Wu, C. C., Noble, W. S., & MacCoss, M. J. (2006). Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries. *Analytical Chemistry*, 78(16), 5678-5684. doi: [10.1021/ac060279n](https://doi.org/10.1021/ac060279n)
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer series in statistics. New York, NY: Springer-Verlag.
- Hertz, H. S., Hites, R. A., & Biemann, K. (1971). Identification of mass spectra by computer-searching a file of known spectra. *Analytical Chemistry*, 43(6), 681-691. doi: [10.1021/ac60301a009](https://doi.org/10.1021/ac60301a009)
- Julian, R. K., Higgs, R. E., Gygi, J. D., & Hilton, M. D. (1998). A method for quantitatively differentiating crude natural extracts using high-performance liquid chromatography-electrospray mass spectrometry. *Analytical Chemistry*, 70(15), 3249-3254. doi: [10.1021/ac971055v](https://doi.org/10.1021/ac971055v)
- Kim, S., Koo, I., Jeong, J., Wu, S., Shi, X., & Zhang, X. (2012). Compound identification using partial and semipartial correlations for gas chromatography-mass spectrometry data. *Analytical Chemistry*, 84(15), 6477-6487. doi: [10.1021/ac301350n](https://doi.org/10.1021/ac301350n)
- Koo, I., Kim, S., & Zhang, X. (2013). Comparative analysis of mass spectral matching-based compound identification in gas chromatography-mass spectrometry. *Journal of Chromatography A*, 1298, 132-138. doi: [10.1016/j.chroma.2013.05.021](https://doi.org/10.1016/j.chroma.2013.05.021)
- Koo, I., Zhang, X., & Kim, S. (2011). Wavelet- and Fourier-transform-based spectrum similarity approaches to compound identification in gas

chromatography/mass spectrometry. *Analytical Chemistry*, 83(14), 5631-5638. doi: 10.1021/ac200740w

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2), 369-411. Available from <http://projecteuclid.org/euclid.ba/1340218343>

Linstrom, P. J., & Mallard, W. G. (Eds.). (2015). *NIST chemistry webbook: NIST standard reference database number 69*. Retrieved from <http://webbook.nist.gov/chemistry/>

Rasmussen, G. T., & Isenhour, T. L. (1979). The evaluation of mass spectral search algorithms. *Journal of Chemical Information and Modeling*, 19(3), 179-186. doi: 10.1021/ci60019a014

Stein, S. E., & Scott, D. R. (1994). Optimization and testing of mass spectral library search algorithms for compound identification. *Journal of the American Society Mass Spectrometry*, 5(9), 859-866. doi: 10.1016/1044-0305(94)87009-8

Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., & Yates, J. R., III. (2003). Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Analytical Chemistry*, 75(10), 2470-2477. doi: 10.1021/ac026424o

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1), 267-288. Available from <http://www.jstor.org/stable/2346178>

Impact of Serial Correlation Misspecification with the Linear Mixed Model

Brandon LeBeau

University of Iowa
Iowa City, IA

Linear mixed models are popular models for use with clustered and longitudinal data due to their ability to model variation at different levels of clustering. A Monte Carlo study was used to explore the impact of assumption violations on the bias of parameter estimates and the empirical type I error rates. Simulated conditions included in this study are: simulated serial correlation structure, fitted serial correlation structure, random effect distribution, cluster sample size, and number of measurement occasions. Results showed that the fixed effects are unbiased, but the random components tend to be overestimated and the empirical Type I error rates tend to be inflated. Implications for applied researchers were discussed.

Keywords: Longitudinal, simulation, linear mixed model

Introduction

Linear mixed models (LMM) have become much more prominent in educational research over the past couple decades, where they are commonly known as hierarchical linear models (HLM) (Raudenbush & Bryk, 2002) or multilevel models (Goldstein, 2010). The mixed portion in the linear mixed model indicates that the model has both fixed and random effects present in the model. These models have become more widely used for a couple of reasons: 1) the advancements in computing which allow for easier and quicker estimation, 2) the notice of the need to model the hierarchical or nested nature of the data, and 3) handles unbalanced data/designs well without any additional work. A few common data collection settings in education where LMM are used include: students nested within classrooms or students nested within schools. For some additional examples of how these models are used in education see Bryk and Raudenbush (1987) and Raudenbush (1988).

Dr. LeBeau is an Assistant Professor in the Educational Measurement and Statistics Program. Email him at: brandon-lebeau@uiowa.edu.

Research Problem

In longitudinal studies, the repeated measures for the same person are likely to be more similar due to the fact that the same person is being measured multiple times on the same measurement scale (Littell, Henry, & Ammerman, 1998). The multiple measurements brings about a dependency due to repeated measurements, or alternatively, there is less information available as the measurement occasions within an individual are correlated. This dependency can be accounted for in the LMM by specifying random effects at the cluster level, the level one covariance matrix, or a combination of the two. In most cases, researchers allow the random effects to account for the dependency due to repeated measures and assume that the variance is the same across the observations with no correlation between the observations (e.g. the correlation between observation one and observation two is zero) at level one. This level one structure is often called an independence structure. For certain repeated measures designs, especially when the repeated measures are collected close in time or correlations among the repeated measures do not decay quickly, random effects alone may not adequately account for the dependency due to the repeated measures and a more complex covariance structure at level one may be needed (Browne & Goldstein, 2010; Goldstein, Healy, & Rabash, 1994).

Unfortunately, few simulation studies have looked at these implications (Ferron, Dailey, & Yi, 2002; Kwok, West, & Green, 2007; Murphy & Pituch, 2009) in a LMM framework. The current study looks to add to this literature by exploring possible implications of misspecifying the level one covariance structure using a computer simulation. The primary question of interest will be the extent to which the misspecification of the variance matrix for the repeated measures biases the parameter estimates (and ultimately inferences as well) for the fixed and random portion of the LMM. Interactions to other assumption violations will also be explored.

The Model

A basic linear mixed model can be written as follows:

$$\mathbf{Y}_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{b}_j + \mathbf{e}_{ij} \quad (1)$$

In this model, the $\mathbf{Y}_{ij}$ is the response variable for the i^{th} level 1 unit nested within the j^{th} level 2 unit. Next is the $\mathbf{X}_{ij}$, which is an $n_i \times p$ matrix of covariates in the model (also known as the design matrix) where n_i is the total number of observations for every individual and p is the number of covariates. This matrix

includes covariates at both level 1 and level 2 as well as covariates that are aggregated over the level 1 units. The $\boldsymbol{\beta}$ in the model is a $p \times 1$ vector representing the fixed effects. Next is the $\mathbf{Z}_{ij}$ which is the design matrix for the random effects. This term is commonly formed from a subset of the columns of $\mathbf{X}_{ij}$. The $\mathbf{b}_j$ are the random effects and are unique for each level 2 unit but are the same for each level 1 unit within a given level 2 unit. The random effects represent the deviation of the j^{th} subject from the group or average growth curve. Finally, the $\mathbf{e}_{ij}$ are the level 1 residuals (i.e. measurement or sampling error) similar to simple linear regression. These represent deviations from the individual growth curves.

This model can also be expressed in matrix form:

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \mathbf{e}_j \quad (2)$$

Model Assumptions

Just like any statistical model, there are model assumptions that need to be satisfied (at least approximately) in order for parameter estimates and inferences to be unbiased. The model assumptions for the LMM are as follows (Raudenbush & Bryk, 2002):

1. The random effects $\mathbf{b}_j$ are independent across level 2 units, normally distributed (multivariate normal when more than one random effect is in the model), and each has a mean of 0 and a covariance matrix $\mathbf{G}$. This can be succinctly written as: $\mathbf{b}_j \sim \text{iid } N(0, \mathbf{G})$.
2. Each of the $\mathbf{e}_{ij}$ are independent and follow a normal distribution with mean 0 and variance σ^2 for every level 1 unit within level two. This can be summed up as: $\mathbf{e}_{ij} \sim \text{iid } N(0, \sigma^2)$.
3. The $\mathbf{e}_{ij}$ are independent of the random effects.

The models considered in this paper are assumed to have a continuous response variable with at least an interval scale of measurement and the within individual errors (i.e. level one errors) are assumed to be approximately normally distributed.

Violation of Model Assumptions

Simulation studies that have data conditions similar to longitudinal data have found little evidence of parameter bias in the fixed or random effects when the random effect distributions are non-normal. However, these studies have reported

confidence intervals for the variance of the random effects with poor coverage when the random effect distributions are not normal, specifically chi-square with one degree of freedom and Laplace distributions (Maas & Hox, 2004a; 2004b). This suggests that the standard errors are underestimated for the variance components of the random effects.

Sample size considerations for the LMM is an important consideration when planning a study. This is especially true since maximum likelihood is asymptotic and require large sample sizes for proper estimation (Maas & Hox, 2004a). Typically, the highest level sample size is of most concern as there are fewer numbers at this level (Maas & Hox, 2004a). This issue is commonly exacerbated for longitudinal studies as the level 1 sample size tends to also be small (i.e. few observations per subject); where 10 observations per subject is considered large (Snijders & Bosker, 1993). Unfortunately, there have been few studies that have studied small level 1 sample sizes commonly found in longitudinal studies.

Simulation studies that have looked at the sample size needed for unbiased estimates for the parameters in general have not found any problems with estimating the fixed effects at level 1 or level 2 (Maas & Hox, 2004a; 2005b; 2005; Browne & Draper, 2000). Additionally, the standard errors for the fixed effects are generally estimated accurately with at least 30 groups (Maas & Hox, 2004a; 2005).

Covariance Structures

The variance structure for the response variable is an important aspect of the LMM; this is where the dependency due to the repeated measures is taken into account. The equation for the variance of the response variable is

$$\text{Var}(\mathbf{Y}_j) = \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_j(\theta) = \mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^T + \sigma_e^2 \mathbf{I}_{n_{1j}} \quad (3)$$

As can be seen from the above equation, the variance is composed of two portions, $\mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^T$ is the portion of the variance that is accounted for by the random effects and the $\sigma_e^2 \mathbf{I}_{n_{1j}}$ is the portion that is accounted for by the level 1 error.

Commonly, researchers choose a simple level 1 error structure. The most common structure specified by researchers assumes homogeneity of variance with no correlation between the time points, known as the independence structure. An example of such a matrix with four time points is as follows:

$$\begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ 0 & 0 & \sigma_e^2 & 0 \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix} \quad (4)$$

where σ_e^2 represents a common variance across the four time points.

Complex variance structures can be achieved by including multiple random effects (e.g. random effects for intercept, time, time2, etc.) and specifying a complex level one error structure. For example, if a researcher fits a model with a random effect for intercept and an independence level one error structure. The covariance structure for the model would look as follows (assuming four time points):

$$\begin{bmatrix} \sigma_e^2 + g_{11} & g_{11} & g_{11} & g_{11} \\ g_{11} & \sigma_e^2 + g_{11} & g_{11} & g_{11} \\ g_{11} & g_{11} & \sigma_e^2 + g_{11} & g_{11} \\ g_{11} & g_{11} & g_{11} & \sigma_e^2 + g_{11} \end{bmatrix} \quad (5)$$

Here σ_e^2 represents the error variance and g_{11} represents the variance of the random intercepts. As can be seen from (5) above when a random intercept is included in the model and an independence structure is assumed at level one, the covariance structure follows a compound symmetry structure (which is what is assumed by RM-ANOVA). Although this structure is not very complex and likely not justifiable for many longitudinal studies, adding more random effects (i.e. a random effect for time) or specifying a more complicated level one error structure (e.g. first order autoregressive, toeplitz, etc.) would produce a more complex covariance structure.

With the inclusion of more complicated error terms, it can be helpful to include additional notation for the level one residual to separate random error and serial correlation denoted as $e_j = e_{(1)j} + e_{(2)j}$. Here $e_{(1)j}$ represents random error and $e_{(2)j}$ represents serial correlation. Serial correlation can be thought of as a random process of an observed profile within an individual that usually decreases as the time lag increases (Diggle, 2002). More simply, serial correlation represents the correlation between two observations on the same individual that depends solely on the time lag between the observations. Explicitly showing the serial correlation

and random error separately in the variance of the response variable leads to the following expression:

$$\text{Var}(\mathbf{Y}_j) = \Sigma_j = \Sigma_j(\theta) = \mathbf{Z}_j \mathbf{G} \mathbf{Z}_j^T + \sigma_e^2 \mathbf{I}_{n_{1j}} + \tau^2 \mathbf{H}_j \quad (6)$$

Different from (3) above, serial correlation is explicitly shown as $\tau^2 \mathbf{H}_j$, where $\mathbf{H}_j$ is an $n_j \times n_j$ matrix where the $(j, k)^{\text{th}}$ element is the correlation between two time points within an individual.

Most researchers when using a LMM tend to assume the level one residual structure follows an independence structure without taking into account the type of data (i.e. cross sectional or longitudinal data). This may be chosen due to the parsimonious nature of the independence model or the researcher believes that including more random effects adequately accounts for the dependency due to repeated measures. However, the following question must be asked, after removing the variation due to the random effects are the within individual residuals independent from one another within an individual (Browne & Goldstein, 2010)? In other words, conditional on the random effects, is it tenable to assume that the within individual residuals are independent? This assumption may not hold in some data situations, especially if the time between observations is very short (i.e. daily or weekly observations) or if the correlation between observations does not decrease very quickly (Browne & Goldstein, 2010; Goldstein et al., 1994). If the level one residuals are not independent of one another, then the level one structure takes a form similar to time series models. See Box and Jenkins (1976) to explore time series models.

Misspecification of the Covariance Structure

There was quite a bit of interest earlier in the history of the LMM on adequately modeling the covariance structure (Chi & Reinsel, 1989; Diggle, 1988; Goldstein et al., 1994; Keselman, Algina, Kowalchuk, & Wolfinger, 1998; 1999; Núñez-Antón & Zimmerman, 2000; Wolfinger, 1996). However, only recently have simulation studies started exploring the impact of misspecification of the level one residual structure (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009). Kwok et al. (2007) defined three useful terms to use when talking about misspecification of the covariance structure: underspecified, overspecified, and general-misspecification. An underspecified covariance structure (US) occurs when the specified matrix is simpler but nested within the true covariance matrix (e.g. compound symmetry is chosen but the true structure is AR(1)). An

overspecified covariance structure (OS) occurs when the specified matrix is more complex than the true covariance matrix but the true covariance matrix is nested within the specified matrix (e.g. ARMA(1, 1) structure chosen but AR(1) is the true structure). Lastly, general-misspecification (GS) occurs when the specified and true covariance matrices are not nested (e.g. TOEP(2) structure chosen but AR(1) is the true structure).

Simulation studies have found little to no bias for fixed effect estimates, however there is evidence of bias in the estimates for the standard errors of the fixed effects (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009). When the covariance structure was US or GS the standard errors for the within-individual intercept and slope were overestimated (Kwok et al., 2007). Not surprisingly, the bias in the variance components can be quite substantial when the covariance structure is ignored. If the covariance structure was US or GS $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$ were overestimated (Ferron et al., 2002; Kwok et al., 2007); OS covariance structures produced the smallest estimates for $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$ (Kwok et al., 2007). As a result of the overestimated $\hat{\tau}_{00}$ and $\hat{\tau}_{11}$, $\hat{\sigma}^2$ tended to be underestimated to compensate (Ferron et al., 2002). Murphy and Pituch (2009) even found that the variance components are biased even when the correct covariance structure was modeled.

These results produced the following general guidelines: if the researcher is only interested in estimates of the fixed effects (i.e. group level estimates) then the researcher may not need to model the covariance structure. However, if the researcher is interested in the variance components, individual growth curves, inferential statistics, or model predictions the researcher should explore alternative structures for the level one covariance structure (Ferron et al., 2002; Kwok et al., 2007; Verbeke & Molenberghs, 2000).

Selecting a Covariance Structure

In most cases when researchers use a LMM, they are interested in doing more than just looking at the fixed effect estimates and some care should be taken to select a covariance structure. However there are no strong descriptive or hypothesis testing procedures to detect serial correlation. The few studies that have explored methods of selecting and detecting serial correlation have found it difficult to empirically select the correct structure (Ferron et al., 2002; Keselman et al., 1998). Another study by Verbeke and Molenberghs (2000) showed that including the serial correlation regardless if it is correctly modeled, is more important than correctly modeling the serial correlation.

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

There are alternative criteria that can be used for selecting the best covariance structure based on the data, these are: Akaike's Information Criterion (AIC), Schwartz's Bayesian Criterion (SBC), or a likelihood ratio test (LRT). Ferron et al. (2002) found that the AIC on average identified the correct structure about 79% of the time. The SBC and LRT identified the correct model less frequently, on average 66% and 71% of the time respectively. However, the variability in correct identification was very large, the AIC ranged from 7% to 100%. Increasing the number of time points, increasing the sample size, and higher levels of autocorrelation improved correct identification (Ferron et al., 2002). In contrast to Ferron et al. (2002), Keselman et al. (1998) found that the AIC or SBC were only able to correctly identify the covariance structure 47% and 35% of the time respectively. The large variability and conflicting results leaves uncertainty in how the researcher should proceed when they desire a test to help decide if serial correlation is present and should be modeled.

Methodology

A factorial research design was used for the computer simulation study. Previous simulation work (Ferron et al., 2002; Kwok et al., 2007; Murphy & Pituch, 2009) have assessed covariance misspecification under perfect model conditions (i.e. normally distributed random effects and residuals); however, a classic study by Micceri (1989), showed that real world data are rarely normally distributed and can deviate quite substantially from a normal distribution. Therefore, simulating conditions more representative of real world data can help inform researchers to the robustness of the estimation algorithm, specifically under small sample size conditions. In addition, missing data tends to be the rule rather than the exception for longitudinal data where the likelihood of missing data commonly increases as time increases (i.e. more likely to encounter more missing data further along in the study). Understanding the implications of covariance misspecification under more common real world data conditions would be helpful and this simulation attempts to inform this area.

In order to simulate conditions that are common in real world data and improve external validity but yet keep the simulation design manageable, the following data conditions were manipulated: the covariance structure (five levels: ID, AR(1), MA(1), MA(2), ARMA(1, 1)), the random effect distribution (three levels: Normal, Laplace, Chi-Square(1)), number of subjects (two levels: 25, 50), and the number of measurement occasions (two levels: 6, 8). This leaves a total of $5 \times 3 \times 2 \times 2 = 60$ simulated data conditions. To avoid finding a single extreme data

condition, five hundred replications were generated for each simulated data condition resulting in $60 \times 500 = 30,000$ total datasets. Statistics were averaged across the 500 replications within each of the 60 simulation conditions. For each dataset, all five of the covariance structures were fitted (i.e. ID, AR(1), MA(1), MA(2), ARMA(1, 1)), resulting in a total of $30,000 \times 5 = 150,000$ models.

Data

Population parameters were generated from data collected by the Minnesota Mathematics Achievement Project (MNMMap). The MNMMap project collected data exploring the relationship between high school mathematics curriculum and subsequent college mathematics grades and course taking for students graduating from a high school in an upper Midwestern state. A retrospective cohort design was used in collecting the data from three sources: high schools, universities or colleges, and the state. The resulting dataset contained student, high school, and college information on more than 20,000 students, from about 300 high schools, and approximately 35 two and four year colleges or universities. In this model, student semester GPA from a college mathematics course will serve as the dependent variable. Time was the primary within-subject variable, ACT score will serve as the single continuous student level predictor and difficulty of the college mathematics course will serve as a time varying covariate. The intercepts and the slope for time were allowed to vary for every student (i.e. a random intercept and a random slope for time were specified in the model). Additional information about the data collection procedures from the MNMMap project can be seen in Harwell et al. (2009) and Post et al. (2010).

Data were simulated according to the following model:

$$Y_{ij} = \beta_0 + \beta_1 \text{time}_{ij} + \beta_2 \text{diff}_{ij} + \beta_3 \text{ACT}_j + \beta_4 \text{ACT}_j : \text{time}_{ij} + b_{0j} + b_{1j} \text{time}_{ij} + e_{(1)ij} + e_{(2)ij} \quad (7)$$

In this equation, let i represent repeated measurements and j represent individuals. The fixed effects are represented by β_0 , β_1 , β_2 , β_3 , and β_4 , time_{ij} represents the within subject time metric, diff_{ij} is a within subject time varying covariate representing the difficulty of the mathematics course, and ACT_j is a continuous subject level covariate representing the mathematics ACT score for each subject. The random components of the model are represented by b_{0j} , b_{1j} , $e_{(1)ij}$,

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Table 1. Parameter values for all terms

Parameter	Value
β_0	2.639
β_1	-0.014
β_2	-0.187
β_3	0.095
β_4	0.003
Var b_{0j}	0.552
Var b_{1j}	0.015
Var e_{ij}	0.549
Φ_1	0.450
θ_1	0.500
θ_2	0.300
Var diff_{ij}	1.250
Var ACT_{1j}	4.905

Note: Var – Variance

and $e_{(2)ij}$ which represent subject specific deviations from the average intercept and slope, deviations from the subject specific growth curves, and serial correlation respectively. Data were simulated from the model shown in (7), where the $e_{(2)ij}$ and the distribution of the random components were the primary differences between the simulated data.

Table 1 shows the population values used to generate the data according to (7). Table 1 reveals that many parameter values are quite small and are reflective of the scale of the dependent variable ranging from zero to four. Of particular note are the small values for β_1 , β_4 , and Var b_{1j} representing the slope for time, the interaction between time and mathematics ACT score, and lastly the variance of the random slopes for time. These small values will have to be kept in mind later as the bias statistic chosen divides by the parameter value.

Analysis

Model convergence, relative bias, and type I error rates were generated for all 150,000 models fitted. Relative bias was computed for all of the fixed effects and the variance components. The formula for relative bias took the form of:

$$\text{Rel. Bias} = \frac{\hat{\theta} - \theta}{\theta} \quad (8)$$

where $\hat{\theta}$ is the parameter estimate (i.e. β_k or $\text{Var}(b_{lj})$) and θ is the parameter value set in the simulation.

The Type I error rate was computed as the proportion of significant fixed effect estimates out of the total number of replications. That is, a Wald test statistic was set up of the form:

$$Z = \frac{\hat{\beta} - \beta}{\text{SE}} \quad (9)$$

where $\hat{\beta}$ is the parameter estimate, β is the simulated parameter value shown in Table 1, and SE is the empirical standard error calculated from the model fit. The Wald test statistic was assumed to follow a standard normal distribution. If there is no bias and the type I error rate is accurate, approximately 5% of the parameter estimates should fall outside of ± 1.96 quantile of the standard normal distribution.

Since a simulation is similar to a completely randomized experiment, the relative bias and type I error rates served as dependent variables and the simulated conditions were treated as independent variables or factors. These variables were analyzed descriptively and inferentially to answer the research questions depicted above.

Inferential Analyses

All of the simulation factors are between-subject factors except for the covariance structure factor which was a within-subject factor as all five covariance structures were fitted to each simulated dataset. Due to the within-subject factor, repeated measures analysis of variance (RM-ANOVA) is a common analysis for this type of data. However, the RM-ANOVA procedure can make interpretation more difficult and increase the burden during estimation. Another data analysis option was to treat all the design factors as between-subject factors and use univariate analysis of variance (UANOVA) to estimate the effects. The UANOVA procedure has the disadvantage of reduced power of the within-subject and mixed interaction effects (i.e. the interaction between the within-subject and between-subject effects). However, with a large sample size in the study ($30,000 \times 5 = 150,000$ total cases in the main analysis) statistical power was not deemed an issue and the UANOVA model was fitted to ease interpretation. A similar analysis was done by Kwok et al. (2007) in their article addressing misspecification of the covariance structure.

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

The initial UANOVA model that was fitted to the relative bias data took the following structure:

$$\begin{aligned}
 Y_{ijklmn} = & \mu + \alpha_{A(j)} + \alpha_{B(k)} + \alpha_{C(l)} + \alpha_{D(m)} + \alpha_{E(n)} + \alpha_{AB(jk)} + \alpha_{AC(jl)} + \alpha_{AD(jm)} \\
 & + \alpha_{AE(jn)} + \alpha_{BC(kl)} + \alpha_{BD(km)} + \alpha_{BE(kn)} + \alpha_{CD(lm)} + \alpha_{CE(ln)} + \alpha_{DE(mn)} \\
 & + \alpha_{ABC(jkl)} + \alpha_{ABD(jkm)} + \alpha_{ABE(jkn)} + \alpha_{ACD(jlm)} + \alpha_{ACE(jln)} + \alpha_{ADE(jmn)} \\
 & + \alpha_{BCD(klm)} + \alpha_{BCE(kln)} + \alpha_{BDE(kmn)} + \alpha_{CDE(lmn)} + \alpha_{ABCD(jklm)} + \alpha_{ABCE(jkln)} \\
 & + \alpha_{ACDE(jlmn)} + \alpha_{BCDE(klmn)} + \alpha_{ABCDE(jklmn)} + e_{ijklmn}
 \end{aligned} \tag{10}$$

The above equation represents a factorial UANOVA that fits all possible interactions. In (10), the α represent cell means, μ is the grand mean, the first set of subscripts, A, B, C, D , and E , represent the five simulation conditions, the subscripts in parentheses, j, k, l, m , and n , index the factor categories, and i depicts the observation number. The model for the empirical type I error rates is simplified compared to (10) because there was only one observation per cell. As a result, all four and five-way interactions were pooled into the error term.

Lastly, significance tests were not used due to the large sample size and statistical power. Instead, effects sizes were computed to determine which factors explained the most variation in the dependent variable. An η^2 statistic was used as the effect size in this analysis and took the following form:

$$\eta^2 = \frac{SS_{\text{trt}}}{SS_{\text{total}}} \tag{11}$$

In the above equation, SS_{trt} is the amount of variation attributable to the treatment of interest (e.g. covariance structure) and SS_{total} is the total sum of squares or the total amount of variation in the dependent variable. η^2 values greater than .001 and .01 were deemed important predictors for the relative bias and empirical type I error rates respectively.

Software

Data generation, model fitting, and analyses were conducted with R (R Development Core Team, 2010). Data generation was undertaken via an author written program. In order to replicate the results, a random seed was chosen and to ensure independent replications, the random number generation was based on the

procedure by L'Ecuyer (L'Ecuyer, Simard, Chen, & Kelton, 2002). This procedure has the advantage of producing very large strings of random numbers without worrying about duplication and supports multiple threads of random number generation which allowed multiple cores of the processor to be used simultaneously improving the data simulation speed. Model fitting was done with the nlme package found in R (Pinheiro, Bates, DebRoy, & Sarkar, 2012). Lastly, in order to check the simulated data conditions, the sample autocorrelation function was plotted to see if the values approximately followed the theoretical autocorrelation function. In addition, the empirical skewness and kurtosis of the simulated random effect distribution was computed to check for accurate random effect simulation. No significant deviations were found.

Results

The convergence rates for study one can be seen in Table 2. This table breaks down the convergence rate of the estimation algorithm by the generated and fitted serial correlation structures. As can be seen from the table, convergence rates tended to be low ranging from a low of 41.6% to a high of 95.9%. Low convergence rates tended to occur when the serial correlation structure was overspecified (e.g. ARMA(1, 1) structure fitted to an AR(1) structure) or when a generally misspecified serial correlation structure was fitted (e.g. AR(1) structure fitted to a MA(1) structure). In general, the AR(1) and ARMA(1, 1) fitted structures had the worst convergence rate compared to the other fitted structures and the independent structure had the best convergence rate, which is not surprising as no additional terms were needed to be estimated with an independent structure.

Relative Bias

Summary statistics for the relative bias of the fixed effects can be seen in Table 3. The table shows that although the mean and median for all of the parameters were very close to zero, the slope terms (i.e. β_1 and β_4) had large amounts of variation as shown by the variance in Table 3. The large amount of variation in the relative bias for those two terms is likely attributable to the small parameter values as seen in Table 1 (i.e. to get the relative bias, the absolute bias was divided by the parameter value which are small for β_1 and β_4).

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Table 2. Convergence rates by generated and fitted serial correlation structure

Gen SC	Fit SC	Convergence %
Ind	Ind	72.48
Ind	AR(1)	68.38
Ind	MA(1)	71.02
Ind	MA(2)	67.23
Ind	ARMA(1, 1)	65.10
AR(1)	Ind	93.88
AR(1)	AR(1)	64.88
AR(1)	MA(1)	81.37
AR(1)	MA(2)	70.78
AR(1)	ARMA(1, 1)	60.45
MA(1)	Ind	92.23
MA(1)	AR(1)	55.12
MA(1)	MA(1)	69.15
MA(1)	MA(2)	65.93
MA(1)	ARMA(1, 1)	63.68
MA(2)	Ind	95.62
MA(2)	AR(1)	61.98
MA(2)	MA(1)	84.50
MA(2)	MA(2)	68.83
MA(2)	ARMA(1, 1)	54.88
ARMA(1, 1)	Ind	98.37
ARMA(1, 1)	AR(1)	42.17
ARMA(1, 1)	MA(1)	88.02
ARMA(1, 1)	MA(2)	72.90
ARMA(1, 1)	ARMA(1, 1)	63.60

Note: Gen – generated, SC – serial correlation, Fit – fitted

Table 3. Summary statistics for relative bias of fixed effects

Term	Mean	Var	Med	Min	Max
β_0	0.0005	0.0054	0.0004	-0.3581	0.4424
β_1	0.0606	26.6853	0.1011	-26.8454	25.1670
β_2	0.0010	0.0905	0.0010	-1.5945	1.7359
β_3	-0.0016	0.1882	-0.0025	-2.4923	2.4803
β_4	0.0579	24.6815	0.0357	-28.2912	30.8497

Note: Var – variance, Med – median, Min – minimum, Max – maximum

The variation in the relative bias for the parameters was explored using ANOVA. No four or five-way interactions had $\hat{\eta}^2$ greater than .001 and were dropped from the models, however all two and three-way interactions were retained.

The results of these final ANOVAs and the resulting $\hat{\eta}^2$ can be seen in Table 4 for all five fixed effect parameters and the variance of the random components. The values in bold in the table are $\hat{\eta}^2$ statistics that are larger than .001.

Looking at the first five columns of Table 4 reveals there are no large $\hat{\eta}^2$ statistics for any of the fixed effects. This means that the simulation conditions do not explain a significant amount of variation in the relative bias of the fixed effects. This suggests that the grand mean relative bias for each of the fixed effects acts as an adequate summary measure for each fixed effect and can be seen in Table 3. Exploring the simple averages shows that relative bias for the two slope terms (i.e. β_1 and β_4) have the largest bias statistics. Even though the slope terms showed slight evidence of bias (.0606 and .0579 for β_1 and β_4 respectively), the relative bias statistic is quite small and would likely not seriously distort any findings.

Summary statistics for the relative bias of the random components can be seen in Table 5. The table shows that on average the variance of the random components tends to be biased and there was significant variation in the relative bias statistics for each term. Since variances can only be positive, it is not surprising that the minimum relative bias is small (approximately -1) compared to the maximum relative bias (approximately 10, 35, and 6.6 for variance of intercept, slope, and within cluster residuals respectively).

The variation in the relative bias statistics for the random components were explored with an ANOVA and the $\hat{\eta}^2$ can be seen in the last three columns of Table 4. These columns reveal that there are variables that explain variation in the relative bias of the random components (i.e. $\hat{\eta}^2 > 0.001$). The strongest effects were the simulated conditions related to the generated and fitted serial correlation structure.

The significant interaction between the generated and fitted serial correlation structures for the random effects are explored in Figure 1. These figures show that fitting an underspecified independence structure has severe consequences in terms of relative bias of the variance of the random effects. More specifically, when an AR(1), MA(1), MA(2), or ARMA(1, 1) structure underlie the data, the independence serial correlation structure produces significantly greater bias compared to fitting other serial correlation structures. For example, when an ARMA(1, 1) structure underlies the data and the serial correlation structure is underspecified as independent, the variance of the intercept and slope are overspecified by over 1.5 times and at least 6 times respectively.

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Table 4. Eta-squared statistics for all terms from ANOVA models

Variable	$\hat{\eta}^2 \beta_0$	$\hat{\eta}^2 \beta_1$	$\hat{\eta}^2 \beta_2$	$\hat{\eta}^2 \beta_3$	$\hat{\eta}^2 \beta_4$	$\hat{\eta}^2 \text{Var } b_0$	$\hat{\eta}^2 \text{Var } b_1$	$\hat{\eta}^2 \text{Var Res}$
<i>N</i>	0.0000	0.0000	0.0001	0.0000	0.0002	0.0023	0.0123	0.0014
<i>p</i>	0.0001	0.0000	0.0000	0.0001	0.0001	0.0010	0.0136	0.0031
RE Dist	0.0001	0.0004	0.0000	0.0001	0.0001	0.0000	0.0000	0.0000
Gen SC	0.0006	0.0008	0.0003	0.0003	0.0001	0.0937	0.0930	0.1704
Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0904	0.0862	0.1984
<i>N:p</i>	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0004	0.0001
<i>N</i> :RE Dist	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0001	0.0000
<i>N</i> :Gen SC	0.0002	0.0001	0.0003	0.0001	0.0001	0.0002	0.0006	0.0003
<i>N</i> :Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001	0.0001	0.0004
<i>p</i> :RE Dist	0.0002	0.0000	0.0001	0.0001	0.0001	0.0001	0.0001	0.0000
<i>p</i> :Gen SC	0.0004	0.0004	0.0001	0.0004	0.0000	0.0002	0.0013	0.0005
<i>p</i> :Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0007	0.0001
RE Dist: Gen SC	0.0002	0.0003	0.0003	0.0004	0.0001	0.0002	0.0003	0.0001
RE Dist: Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Gen SC:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0670	0.0548	0.1658
<i>N:p</i> :RE Dist	0.0001	0.0000	0.0001	0.0001	0.0001	0.0002	0.0002	0.0001
<i>N:p</i> :Gen SC	0.0002	0.0000	0.0002	0.0002	0.0001	0.0001	0.0002	0.0000
<i>N:p</i> :Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0001
<i>N</i> :RE Dist:Gen SC	0.0002	0.0004	0.0001	0.0004	0.0006	0.0001	0.0002	0.0003
<i>N</i> :RE Dist:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>N</i> :Gen SC:Fit SC	0.0000	0.0001	0.0000	0.0000	0.0000	0.0006	0.0002	0.0019
<i>p</i> :RE Dist:Gen SC	0.0002	0.0001	0.0004	0.0001	0.0003	0.0003	0.0004	0.0001
<i>p</i> :RE Dist:Fit SC	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
<i>p</i> :Gen SC:Fit SC	0.0000	0.0000	0.0000	0.0001	0.0000	0.0001	0.0006	0.0005
RE Dist:Gen SC:Fit SC	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001

Note: Bold numbers are > 0.001, *N* – cluster sample size, *p* – within cluster sample size, Gen – generated, RE Dist – random effects distribution, SC – serial correlation, Fit – fitted, “:” represents an interaction

Table 5. Summary statistics for relative bias of random components

Term	Mean	Var	Med	Min	Max
$\hat{\eta}^2 \text{Var } b_{0j}$	0.4012	0.6942	0.2904	-1.0000	10.0186
$\hat{\eta}^2 \text{Var } b_{1j}$	1.9116	9.2561	1.1211	-1.0000	35.4700
$\hat{\eta}^2 \text{Var Res}$	0.1222	0.2645	-0.0151	-0.7943	6.6436

Note: Var – variance, Med – median, Min – minimum, Max – maximum

The AR(1) and ARMA(1, 1) fitted structures tend have the smallest bias statistics for the variance of the random effects compared to the other structures, which may suggest that the moving average component does not aid in modeling

serial correlation in longitudinal data. Lastly, even when the correct structure is modeled there is still evidence of bias in the variance of the random effects and in many cases the correct fitted structure does not produce the smallest average relative bias statistics.

Lastly, [Figure 2](#) shows that the variance of the residuals tend to be underestimated when an underspecified independence structure is fit, however this underestimation is not as large as the overspecification found in the random effects. The largest amount of bias occurs when the underlying structure is ARMA(1, 1), which tends to produce average relative bias statistics for the residuals that are comparable to the average relative bias for the variance of the intercept. Except for the systematic underestimation when an independence structure was fitted when serial correlation was present, the average relative bias still tends to be positive suggesting that all of the random components are overestimated when serial correlation is present.

Type I Error Rate

Even though there was no evidence of bias in the fixed effects under any of the simulated data conditions, the random components did show evidence of bias; therefore, the standard errors of the fixed effects may not be accurate. This may cause the type I error rate to be too conservative (type I error rate smaller than the specified α) or too liberal (type I error rate greater than the specified α).

Box plots can be seen in [Figure 3](#) and show the empirical type I error rates for each of the fixed effect parameters. This figure shows that the median empirical type I error rate for the fixed effects tends to be slightly above the expected $\alpha = 0.05$, however β_0 and β_3 both include 0.05 in the middle 50% of the distribution. β_0 , β_1 , and β_4 have median type I error rates around 0.06, whereas β_2 has a median around 0.07. The variability in the five box plots tend to be similar indicated by the size of the interquartile range. Since there does appear to be variability in the empirical type I error rates, these will be modeled inferentially. [Table 6](#) shows the $\hat{\eta}^2$ statistics for the empirical type I error rates for all terms up to three-way interactions. All higher order interaction terms were pooled into the error.

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Figure 1. Relative bias of random effects by generated and fitted serial correlation structure; variance of b_{0j} (left) and b_{1j} (right)

Figure 2. Relative bias of the variance of the residuals by generated and fitted serial correlation structures

As can be seen from the table there were numerous effect sizes greater than 0.01. Some of the largest effects were the cluster sample size, the interaction between the generated serial correlation structure and random effect distribution, and the three way interactions between the generated serial correlation structure, the random effect distribution, and the cluster sample size or the within cluster sample size. These large effects were around 0.10 suggesting that approximately 10% of the variation in the type I error rates can be explained by each of these terms.

The average empirical type I error rate for β_0 by the generated serial correlation structure, random effect distribution and the cluster sample size can be seen in Figure 4. From the figure, cluster sample sizes of 25 tend to have larger average type I error rates compared to cluster sample sizes of 50. There also was a lot of variability in the average type I error rate as the generated serial correlation structure differs, with the AR(1) structure having the smallest amount of variation. The empirical type I error rate was the smallest when the simulated random effect distribution was normally distributed.

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Figure 3. Box plot of type I error rates by parameter

Lastly, the scale of the y -axis should be taken into account. Although there is variability in the average type I error rates, this variability ranges from about 0.04 to just over 0.07 with an even smaller range when the cluster size is 50. Even though most conditions are inflated, they may not be inflated enough to significantly concern applied researchers.

Patterns for the empirical type I error rates were similar for the other parameters (i.e. $\beta_1, \dots, \beta_4$) and are not presented graphically. In addition, the patterns were also similar for the three way interaction between the generated serial correlation structure, random effect distribution, and within cluster sample size and these graphs are not presented. The range of possible average empirical type I error rates were smaller for this second three way interaction compared to the one shown in Figure 4.

Table 6. Eta-squared statistics for all terms from ANOVA models

Variable	$\hat{\eta}^2 \beta_0$	$\hat{\eta}^2 \beta_1$	$\hat{\eta}^2 \beta_2$	$\hat{\eta}^2 \beta_3$	$\hat{\eta}^2 \beta_4$
<i>N</i>	0.0108	0.1111	0.1014	0.0150	0.0866
<i>p</i>	0.0037	0.0005	0.0152	0.0000	0.0065
RE Dist	0.1133	0.0119	0.0617	0.0282	0.0286
Gen SC	0.0416	0.0518	0.0338	0.0196	0.0857
Fit SC	0.0086	0.0145	0.1579	0.0049	0.0137
<i>N:p</i>	0.0476	0.0385	0.0240	0.0129	0.0038
<i>N</i> :RE Dist	0.0160	0.0147	0.0631	0.0072	0.0066
<i>N</i> :Gen SC	0.0300	0.0090	0.0352	0.1305	0.0755
<i>N</i> :Fit SC	0.0037	0.0030	0.0079	0.0017	0.0024
<i>p</i> :RE Dist	0.0102	0.0188	0.0096	0.0075	0.0638
<i>p</i> :Gen SC	0.0468	0.0306	0.0027	0.0581	0.0356
<i>p</i> :Fit SC	0.0030	0.0025	0.0034	0.0131	0.0088
RE Dist: Gen SC	0.0339	0.0525	0.0354	0.0814	0.0820
RE Dist: Fit SC	0.0060	0.0038	0.0043	0.0117	0.0035
Gen SC:Fit SC	0.0151	0.0412	0.0351	0.0180	0.0712
<i>N:p</i> :RE Dist	0.0196	0.0051	0.0047	0.0218	0.0338
<i>N:p</i> :Gen SC	0.1475	0.0156	0.0601	0.0269	0.0338
<i>N:p</i> :Fit SC	0.0010	0.0021	0.0115	0.0012	0.0005
<i>N</i> :RE Dist:Gen SC	0.0397	0.0713	0.0523	0.0747	0.0380
<i>N</i> :RE Dist:Fit SC	0.0070	0.0084	0.0132	0.0188	0.0084
<i>N</i> :Gen SC:Fit SC	0.0128	0.0109	0.0103	0.0191	0.0111
<i>p</i> :RE Dist:Gen SC	0.1112	0.0989	0.0792	0.0969	0.0961
<i>p</i> :RE Dist:Fit SC	0.0023	0.0038	0.0152	0.0099	0.0107
<i>p</i> :Gen SC:Fit SC	0.0067	0.0193	0.0147	0.0254	0.0103
RE Dist:Gen SC:Fit SC	0.0309	0.0205	0.0254	0.0355	0.0228

Note: Bold numbers are > 0.01, *N* – cluster sample size, *p* – within cluster sample size, Gen – generated, RE Dist – random effect distribution, SC – serial correlation, Fit – fitted, “:” represents an interaction

Sensitivity Analysis

An arcsine transformation was done on the empirical type I error rates that were analyzed above. The transformation was performed for two reasons, first to remove the hard 0 and 1 boundaries of the proportion metric, and second to remove the mean and variance relationship of the proportion metric. This transformation took the following form:

SERIAL CORRELATION MISSPECIFICATION WITH THE LMM

Figure 4. Mean type I error rate for β_0 by generated serial correlation structure, random effect distribution, and cluster sample size

$$\begin{aligned}
 \hat{p}'_k &= 2\sqrt{\sin^{-1} \hat{p}_k}, 0 < \hat{p}_k < 1 \\
 \hat{p}'_k &= 2\sqrt{\sin^{-1} \left(\frac{1}{4} R_k \right)}, \hat{p}_k = 0 \\
 \hat{p}'_k &= 3.14 - 2\sqrt{\sin^{-1} \left(\frac{1}{4} R_k \right)}, \hat{p}_k = 1
 \end{aligned} \tag{12}$$

where R refers to the number of simulation replications. After making the transformation, the transformed empirical type I error rates will be normally distributed with mean p'_k and variance $1/R_k$ (Marascuilo & McSweeney, 1977). After the transformation was performed, a similar model was fitted to the data as discussed above except now the average arcsine transformed empirical type I error

rate was used as the dependent variable. Just as before, η^2 served as the effect size to identify variables that explained significant variation in the dependent variable as opposed to p -values.

The effect sizes calculated from the arcsine transformed empirical type I error rates were similar to the model left in the original proportion metric with no additional variables identified as significant. Since the results were similar, interpretations made above in the original proportion metric are similar regardless of the scale of measurement which adds to the robustness of results.

Discussion

The current Monte Carlo study explored the implications for the LMM when model assumptions have not been adequately met. Five different generated serial correlation structures, independent, AR(1), MA(1), MA(2), and ARMA(1, 1) were explored in the current Monte Carlo study along with three different simulated random effect distributions, normal, chi-square (1), and Laplace.

Study results showed that the fixed effects on average were unbiased and none of the simulation conditions explained significant variation in the relative bias of the fixed effects for either of the studies. However, there was evidence of bias in the variance components and simulation conditions did explain significant variation in the average relative bias. This is similar to previous research when serial correlation was not modeled and the random components were normally distributed (Kwok et al., 2007; Murphy & Pituch, 2009).

Unfortunately, no real pattern to which fitted serial correlation is best emerged, for example overspecified or underspecified covariance structures did not consistently provide better estimates of the random components. Instead including some measure of serial correlation, when present, helps to alleviate some bias concern for the random effects. However, even correctly modeling the serial correlation structure tended to produce biased random components of the model. The AR(1) and ARMA(1, 1) tended to produce the smallest amounts of bias in the random components, however the convergence rate was impacted when these additional parameters were included in the model.

For both the fixed effects and random components, the simulated random effect distribution did not explain significant variation in the relative bias statistics. This is contrary to prior work exploring the robustness of the LMM to normality assumptions (Maas & Hox, 2004a; LeBeau, 2013). Results from this prior work found that the simulated random effect distribution did not produce bias in the fixed effects, but did introduce bias into the random effects. However, these studies did

not build explanatory models to see which study conditions explain variation in the relative bias statistics. Adding the more complicated serial correlation structures may have influenced this relationship and overpowered the influence of the non-normal random effect distribution.

This Monte Carlo study also explored the type I error rates of the five fixed effects. The fixed effects were all slightly elevated compared to the $\alpha = 0.05$ level. Increasing the sample size at both levels of the model was the best way to help limit the slight inflation found in the empirical type I error rates. Trends regarding the generated or fitted serial correlation structure and the simulated random effect distribution were not as clear.

Recommendations

Recommendations for researchers come in three different groups. First, if the researcher is only interested in the estimates of the fixed effects, then one does not need to worry about the serial correlation. The results showed that the relative bias for the fixed effects were not affected by any of the simulation conditions studied, including the generated or fitted serial correlation structures, random effect distribution, sample size considerations, or missing a random effect. These results are similar to other Monte Carlo studies with the linear mixed model (Ferron et al., 2002; Kasim & Raudenbush, 1998; Kwok et al., 2007; Maas & Hox, 2004a; Murphy & Pituch, 2009).

However, if the researcher is interested in estimates of the random effects, more care needs to be taken. In general, the random effects tend to be overestimated when serial correlation is present and ignored (i.e. an independence structure is assumed to underlie the data when this is not the case). Although still overestimated, more measurement occasions (i.e. within cluster sample size) and fitting an AR(1) or ARMA(1, 1) serial correlation structure tends to limit the overestimation of the random effects.

Lastly, if the researcher is interested in inference about the fixed effects care needs to be taken to explore whether serial correlation is present in the data. This is especially important when the number of individuals (clusters) and the number of repeated measurements are small. Although not severely inflated, it is likely that the α value specified by researchers is slightly larger in practice.

Unfortunately, there is no a priori test to directly test for the presence of serial correlation in the data. To look for serial correlation, a variogram could be used or descriptively looking at the average correlations between measurement occasions. Another tactic would be to use a procedure such as the likelihood ratio test or model

fit indices such as the AIC or SBC to see if modeling the serial correlation improves model fit. Unfortunately, these methods have not been very reliable in selecting the correct structure (Ferron et al., 2002; Keselman et al., 1998).

Future Work

Future work exploring reasons for the poor convergence rate of the models is needed. Increasing the variances of the random components to see if that aids the poor convergence rates would be helpful. Increasing the variance of the random components may also have an impact on the empirical type I error rates and would be useful to explore.

Detecting serial correlation when present in the data is another area of work that needs to be explored. Currently it is difficult to detect serial correlation from the data putting researchers in a difficult position when searching for serial correlation in their data. Procedures to use when looking for serial correlation in the data would provide guidance for researchers. Exploring additional missing data structures would also be useful. The current study used dropout as a missing data structure as this commonly occurs in longitudinal data, however it is not the only way missing data occurs. For example, having a subject to re-enter the study after missing a measurement occasion is also common in longitudinal data.

Finally, additional work that relaxes the assumption that random effects are uncorrelated across clusters, extending the work done by Browne and Goldstein (2010) in a Bayesian framework, could be a new extension of this group of models. This would give researchers the flexibility of modeling three levels of nesting through the use of a two level model. Situations where this would be most helpful would be when relatively few level three units are sampled, for example when only five schools are sampled. It would likely not be possible to model this third level of nesting with only five units, however accounting for this dependency through correlated random effects at level two may be useful and necessary if the third level of nesting accounts for a significant amount of variation.

References

- Box, G. & Jenkins, G. (1976). *Time series analysis: Forecasting and control*. Eaglewood Cliffs, NJ: Prentice Hall.
- Browne, W. & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, 15(3), 391-420. doi: [10.1007/s001800000041](https://doi.org/10.1007/s001800000041)
- Browne, W. & Goldstein, H. (2010). MCMC sampling for multilevel model with nonindependent residuals within and between cluster units. *Journal of Educational and Behavioral Statistics*, 35(4), 453-473. doi: [10.3102/1076998609359788](https://doi.org/10.3102/1076998609359788)
- Bryk, A. & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101(1), 147-158. doi: [10.1037/0033-2909.101.1.147](https://doi.org/10.1037/0033-2909.101.1.147)
- Chi, E. & Reinsel, G. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, 84(406), 452-459. doi: [10.1080/01621459.1989.10478790](https://doi.org/10.1080/01621459.1989.10478790)
- Diggle, P. (1988). An approach to the analysis of repeated measurements. *Biometrics*, 44(4), 959-971. doi: [10.2307/2531727](https://doi.org/10.2307/2531727)
- Diggle, P. (2002). *Analysis of longitudinal data*. New York, NY: Oxford University Press.
- Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, 37(3), 379-403. doi: [10.1207/S15327906MBR3703_4](https://doi.org/10.1207/S15327906MBR3703_4)
- Goldstein, H. (2010). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.
- Goldstein, H., Healy, M., & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13(16), 1643-1655. doi: [10.1002/sim.4780131605](https://doi.org/10.1002/sim.4780131605)
- Harwell, M., Post, T., Cutler, A., Maeda, Y., Anderson, E., Norman, K., & Medhanie, A. (2009). The preparation of students from national science foundation-funded and commercially developed high school mathematics curricula for their first university mathematics course. *American Educational Research Journal*, 46(1), 203-231. doi: [10.3102/0002831208323368](https://doi.org/10.3102/0002831208323368)
- Kasim, R., & Raudenbush, S. (1998). Application of gibbs sampling to nested variance components models with heterogeneous within-group variance.

Journal of Educational and Behavioral Statistics, 23(2), 93-116. doi: 10.3102/10769986023002093

Keselman, H., Algina, J., Kowalchuk, R., & Wolfinger, R. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Communications in Statistics – Simulation and Computation*, 27(3), 591-604. doi: 10.1080/03610919808813497

Keselman, H., Algina, J., Kowalchuk, R., & Wolfinger, R. (1999). A comparison of recent approaches to the analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, 52(1), 63-78. doi: 10.1348/000711099158964

Kwok, O., West, S., & Green, S. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A Monte Carlo study. *Multivariate Behavioral Research*, 42(3), 557-592. doi: 10.1080/00273170701540537

LeBeau, B. (2013, April). *Impact of non-normal random components on the linear mixed model*. Poster session presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

L'Ecuyer, P., Simard, R., Chen, E., & Kelton, W. (2002). An object-oriented random-number package with many long streams and substreams. *Operations Research*, 50(6), 1073-1075. doi: 10.1287/opre.50.6.1073.358

Littell, R., Henry, P., & Ammerman, C. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*, 76(4), 1216-1231. Retrieved from <https://dl.sciencesocieties.org/publications/jas/abstracts/76/4/1216>

Maas, C. & Hox, J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, 58(2), 127-137. doi: 10.1046/j.0039-0402.2003.00252.x

Maas, C. & Hox, J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, 46(3), 427-440. doi: 10.1016/j.csda.2003.08.006

Maas, C. & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral & Social Sciences*, 1(3), 86-92. doi: 10.1027/1614-2241.1.3.86

Marascuilo, L. & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole Pub. Co.

- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1), 156-166. doi: [10.1037/0033-2909.105.1.156](https://doi.org/10.1037/0033-2909.105.1.156)
- Murphy, D. & Pituch, K. (2009). The performance of multilevel growth curve models under an autoregressive moving average process. *The Journal of Experimental Education*, 77(3), 255-284. doi: [10.3200/JEXE.77.3.255-284](https://doi.org/10.3200/JEXE.77.3.255-284)
- Núñez-Antón, V. & Zimmerman, D. (2000). Modeling nonstationary longitudinal data. *Biometrics*, 56(3), 699-705. doi: [10.1111/j.0006-341X.2000.00699.x](https://doi.org/10.1111/j.0006-341X.2000.00699.x)
- Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2012). The nlme package: Linear and nonlinear mixed effects models (Versions 3.1-103) [Software]. Available from <https://cran.r-project.org/web/packages/nlme/index.html>
- Post, T., Medhanie, A., Harwell, M., Norman, K., Dupuis, D., Muchlinski, T., ... Monson, D. (2010). The impact of prior mathematics achievement on the relationship between high school mathematics curricula and postsecondary mathematics performance, course-taking, and persistence. *Journal for Research in Mathematics Education*, 41(3), 274-308. Available from <http://www.jstor.org/stable/20720139>
- R Development Core Team. (2010). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org>
- Raudenbush, S. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational and Behavioral Statistics*, 13(2), 85-116. doi: [10.3102/10769986013002085](https://doi.org/10.3102/10769986013002085)
- Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.
- Snijders, T. & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational and Behavioral Statistics*, 18(3), 237-259. doi: [10.3102/10769986018003237](https://doi.org/10.3102/10769986018003237)
- Verbeke, G. & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer.
- Wolfinger, R. (1996). Heterogeneous variance-covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, 1(2), 205-230. Available from <http://www.jstor.org/stable/1400366>

Validity of the Persian Blog Attitude Questionnaire: An Evidence-Based Approach

Vahid Aryadoust
National University of Singapore
Singapore

Zahra Shahsavar
Shiraz University of Medical Sciences
Shiraz, Iran

Validity evidence is provided for a Persian blog attitude questionnaire (P-BAQ). P-BAQ was administered to 565 Iranians and factor analysis and rating scale model identified affective, behavioral, and perseverance, and confidence dimensions underlying the data. P-BAQ's validity argument was supported by the theoretical and psychometric evidence, although adding a few items to the instrument would improve its construct representativeness.

Keywords: blog, confirmatory factor analysis, Persian blog attitude questionnaire, principal component analysis, Rasch-Andrich rating scale model, validity argument

---

Introduction

Advanced technology has resulted in a wide range of Web 2.0 tools such as blogs that can be used for various purposes such as disseminating information and launching discussions (Wang & Woo, 2010). Blogs have recently been adapted into educational settings; they have facilitated information access and interaction among students and allowed them to collaborate on group projects, edit or add to page contents, and discuss their projects with their peers in a friendly and less formal environment (Liaw, Huang, Chen, 2007).

Recently, researchers have shown an increased interest in measuring students' attitudes toward technology (Chen, Shih, & Liu, 2013). Blog attitude refers to users' inclination to use blogs for various purposes such as disseminating

*Footnote: For consistency and to avoid confusions, the term dimension is used across PCA, CFA, and RSM analyses.

Vahid Aryadoust is Lecturer at the Centre for English Language Communication of the National University of Singapore. Email him at: vahid.aryadoust@nus.edu.sg. Zahra Shahsavar is Assistant Professor at Shiraz University of Medical Sciences. Email her at: shahsavarzahra@gmail.com.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

information and introducing educational programs. Much of the interest in studying blog attitude is due to the reported influence of attitude on students' educational performance (Ayres, 2002). Studies show that in educational settings where blogs are used, students who harbor a positive attitude toward blogs would find learning attractive, efficient, and practical (Fageeh, 2011; Chu & Kamal, 2008). Positive attitudes toward blogs help students adapt themselves to a new set of learning and teaching techniques that technology offers, thereby allaying their stress and discomfort (Al-Tamimi & Shuib, 2009; McIntosh, 2009). By contrast, students with negative attitudes toward blogs would be less willing to involve themselves in learning because they might view technology as a threat to their learning (Shahsavari & Tan, 2011) or a threat to their well-established and traditional techniques (Pektas & Erkip, 2006).

To assess learners' attitudes toward technology, researchers have employed questionnaires. Substantial research has focused on examining learners' attitudes toward computers by using popular computer attitude scales (e.g., DeVellis, 2003; Loyd & Gressard, 1984; Selwyn, 1997; Shaft, Sharfman, & Wu, 2004). However, relative to computers, little research has been performed on blog attitude (e.g., Blackstone, Spiri, & Naganuma, 2007; Harwood, 2010). One of the limitations of the blog attitude questionnaires is that they are often written in English or are not clearly translated into the respondents' mother tongue. According to Harkness (2008), the inaccuracy of the translated questionnaire may adversely affect respondents' comprehension and consequently undermine the reliability of the data. Instruments should be presented in respondents' own mother tongue to enhance the quality and precision of data.

In addition, most blog attitude questionnaires present no solid evidence of psychometric quality and no explicitly articulated validity argument (see Kane, 2013, for a discussion on the importance of these requirements). To develop psychometrically valid questionnaires, an item pool containing items from available questionnaires should be initially developed. Items adapted from instruments written in a foreign language should be translated into the respondents' mother tongue. To ascertain construct equivalence—i.e. that the items in English and the respondents' mother tongue tap the same construct—the precision of the translated items should be verified by experts (Dörnyei & Taguchi, 2010).

The intent of the present study is to develop and examine the psychometric features of the Persian Blog Attitude Questionnaire (P-BAQ). Developing P-BAQ is primarily motivated by the lack of validated Persian blog attitude questionnaires and a need for further research into Iranian students' blog attitude.

This lack can make it difficult for researchers in Iran (and other Farsi/Persian speaking countries) to perform research on blog attitude and obtain reliable and reproducible results. Although the need for validation and rigorous data analysis has been identified in attitude measurement research (e.g., [Bangert, 2009](#); [Morse, Gullekson, Morris, & Popovich, 2011](#); [Zhang, 2007](#)), previous research tends to overlook the importance of validation as an argument. The researchers survey the relevant literature and use factor analysis and Rasch-Andrich rating scale model (RSM) to examine the psychometric quality of P-BAQ. Subsequently, the evidence gleaned from the psychometric analyses is used to build a coherent validity argument for the instrument.

Literature Review

Attitude

Pickens (2005) defines attitude as learners' way of evaluating objects (or people) positively or negatively. There is consensus among researchers on the effect of students' attitude toward technology tools on their motivation and learning ([Vandewaetere & Desmet, 2009](#)). However, there has yet to be an agreement on the structure of attitude as a measureable construct; this lack of consensus has resulted in the development of multidimensional and unidimensional models of attitude in technology attitude research ([Franzoi, 2003](#)).

Researchers endorsing attitude multidimensionality propose several underlying dimensions, three commonly stated of which are affective, behavioral, and cognitive dimensions ([Eagly & Chaiken, 1993](#); [McLeod, 2009](#); [Hogg & Vaughan, 2011](#); [Mantle-Bromley, 1995](#); [Mantle-Bromley & Miller, 1991](#); [Zimbardo & Leippe, 1991](#)). [McLeod \(2009\)](#) argues that these three dimensions would constitute a basic model for attitude, suggesting that attitude might emerge as a three-dimensional construct. The affective dimension refers to individuals' feelings toward people and using objects in their life. For example, students might express their dislike for course blogs, which results from their negative attitudes toward it. The behavioral dimension refers to individuals' aim to use objects or act in a particular situation. For example, students who dislike course blogs may be less participatory and attentive in class than the students who find blogging enjoyable. The cognitive dimension engages individuals' knowledge and beliefs about objects or people. For example, the students who disapprove of using course blogs may believe that technology is of no practical use for learning and teaching ([McLeod, 2009](#)).

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Some researchers who endorse attitude's multidimensionality have excluded the cognitive dimension, arguing that cognition is a determinant of affective and behavioral dimensions rather than an attitudinal component. For example, Min (1998, p. 23) defines attitude as "an evaluative response to the environment, ideas, objects, and other people" which is predicted by individuals' beliefs, knowledge, and perceptions. Similarly, Mantle-Bromley (1995) argues that attitude is composed of individuals' emotions such as detest, fear, and love as well as behaviors. Selwyn's (1997) application of factor analysis yields empirical evidence backing Min's (1998) and Mantle-Bromley's (1995) conceptualizations of attitude, eliminating the cognitive dimension due to its poor loading coefficients. Tsai, Lin, and Tsai's (2001) and Kay's (1993) studies also yield highly similar results to Selwyn's study, casting doubt on the validity of the postulated cognitive dimension.

Contrary to the studies supporting the multidimensional structure of attitude, some researchers argue that the posited attitude dimensions may be statistically inseparable or unidimensional. Daud (1995) applied factor analysis on a pool of items measuring language teachers' and students' attitude toward computers, where a single attitude factor emerged. Smith, Caputi, and Rawstorne (2000) also found that attitude's posited dimensions were inseparable, arguing that the items that elicit students' attitudinal and "psychological reactions" might not load on distinct factors. Although this line of research is relatively underdeveloped, it is important to compare the fit of unidimensional and multidimensional structures to the data to ascertain the psychometric validity of the scale (Teo, 2012).

Finally, a number of educators have used the basic model of attitude to develop computer and internet attitude questionnaires (e.g., Jones & Clarke, 1994; Liaw, 2002; Fančovičová & Prokop, 2008; Sam, Othman, & Nordin, 2005). However, there is still no blog attitude questionnaire—specifically for Persian (and Farsi) speakers—with a reliable psychometric structure and validity argument. The present study seeks to develop a Persian blog attitude questionnaire (P-BAQ); examine its psychometric features; and lay out a validity argument for it.

Building P-BAQ's Validity Argument

For the validation of psychometric instruments, Kane (2006, 2013) proposed a framework with two main components: i) the claimed uses of the instrument and interpretation of its scores and ii) the supporting evidence. The advantage of this framework is that research findings can be organized into a "cohesive treatment"

for validation of a psychometric instrument (Aryadoust, 2013, p. 12). In greater detail, this framework consists of inferences made from the data, the evidence supporting these inferences, any assumptions being made, and the justifications for these assumptions (Aryadoust, Mehran, & Alizadeh, 2014). Hence, for the purpose of this study, we assert:

- (a) *The domain definition inference* infers that Iranian tertiary students have a multicomponential attitude toward blogs that include an affective component (i.e., positive or negative feelings about the use of blogs for education) and a behavioral component (i.e., a preference for using blogs for education). This inference requires the following assumptions: i) we can determine the multiple components of attitude toward blogs and ii) we can measure these components. As the literature survey indicated that attitude toward blogs has multiple components, the first assumption is supported. The pool of attitude items identified from the literature supports the second assumption. Further discussion of the literature survey supporting the content-representativeness of the items can be found under *Attitude*.
- (b) *The translation inference* occurs during the translation of the items into Persian and infers that the items have been clearly translated and do not contain any ambiguous language. This inference requires the following assumptions: i) the items can be translated and ii) the translated items can be unambiguously understood by Persian speakers. The use of a panel of expert translators to consult on the clarity and accuracy of the translations supports these assumptions.
- (c) *The evaluation inference* infers that participants' responses are consistent and thus produce reliable data. This inference assumes that the scoring categories (i.e., a four-category Likert scale) is suitable. This assumption is warranted because a four-category Likert scale prevents respondents from only endorsing the mid-points of the scale ("flat-lining"). The psychometric qualities of the Likert scale are supported by the monotonic increments of its response categories and the fit statistics for each response category. This is supported by the RSM results, as presented below.
- (d) *The generalization inference* infers that observed scores can be generalized to the universe of scores. This inference assumes that performance on the survey represents performance on the universe of

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

items. This assumption is warranted by psychometric measurements and the reliability statistics estimated by RSM.

- (e) *The explanation inference* infers a relationship between the observed scores and the underlying construct. This inference requires the following assumptions: i) P-BAQ has good construct representativeness, ii) the sub-components of P-BAQ are highly correlated, and iii) there are no construct-irrelevant factors. These assumptions are warranted because the structure of P-BAQ is supported by theory, P-BAQ's sub-components are related and can be distinguished from each other, and the psychometric qualities of P-BAQ are reliable with no construct-irrelevant factors, which can be confirmed by dimensionality analysis. Psychometric analysis further supports this inference; specifically, principal component analysis (PCA) supports the multidimensional structure, confirmatory factor analysis (CFA) supports a multidimensional structure in a different sample, and RSM analysis supports the unidimensionality of P-BAQ's sub-components and shows there are no gaps in the item-person map (Aryadoust, 2013). (For consistency, the term dimension is used across PCA, CFA, and RSM analyses.)

As a one-to-one correspondence between the data analysis techniques and postulations is not possible (Aryadoust, 2013; Kane, 2013), multiple psychometric analyses are performed to obtain data supporting the inferences described above, e.g., RSM results can be used to support the generalization and explanation inferences while CFA results can be used to support the explanation inference.

Methodology

Participants

A sample of 565 Iranian tertiary-level students aged between 17 and 30 ($M = 21.98$; $SD = 2.47$) completed P-BAQ in late 2013. Among these students, 254 (44.6%) were male and 298 (53.1%) were female (missing gender information: 13; 2.3%), thereby yielding a balanced gender distribution. Of these, 361 reported their field of study, as follows: Dentistry: 84 (15%); Medicine, Hygiene, and Speech Therapy: 68 (12%); English Translation: 25 (4.5%); Computer Engineering: 117 (20.9%); and Industrial Engineering: 67 (11.9%). All students indicated their familiarity with blogs; had personal computers and home

Internet access; and had used blogs and/or one of the Web 2.0 tools such as email and Facebook.

The sample was randomly divided into two sub-sets containing 282 and 283 participants for psychometric analysis (see below).

Developing P-BAQ

Before attempting to develop their own tool, existing technology attitude questionnaires were surveyed, such as those developed by Blackstone, Spiri, and Naganuma (2007), Harwood (2010), and Shahsavar and Tan (2012). This resulted in an initial pool of 100 English language items, which was reduced after ambiguous/imprecise items were removed (see Aryadoust et al., 2014, for more details of such a process).

This refined item pool was then submitted to a panel of five experts – two with PhD in applied linguistics or educational/instructional technology and three PhD candidates in applied linguistics or English. The experts were identified as having expert knowledge in incorporating new technology (blogs, social media) in education, as well as practical experience developing questionnaires about online learning. Experts were selected with this background to optimize construct-representativeness. Each expert performed an independent review of item pool, resulting in 29 items identified by at least three of the experts.

These 29 items were then translated into Persian by the authors. The translations were submitted to experienced translators who reviewed the clarity, naturalness, and appropriateness (Farahzad, 1992) of the translation. In the end, a 29-item P-BAQ was formalized that evaluates Iranian blog users' attitudes using a four-point response scale: *strongly disagree* (1), *disagree* (2), *agree* (3), or *strongly agree* (4) (see the Appendix).

Data Analysis

Principal Component Analysis (PCA)

The univariate normality of items estimated by skewness and kurtosis indices were initially examined—values falling between -1 and +1 indicate normality. To determine the constituent structure of P-BAQ, multiple PCA were performed on the first sub-set ($n = 280$) on IBM SPSS Statistics for Windows, Version 21.0 (IBM Corp., 2012).

Because Items 1, 4, and 19 displayed cross-loading patterns, they were eliminated one at a time. To achieve the optimal results, different rotation

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

methods were performed including oblique rotation methods which allow the components to correlate (i.e., direct oblimin and promax) and orthogonal rotation which does not permit the components to correlate (i.e., equimax, quartimax, and varimax).

The Kaiser criteria (where eigenvalues greater than unity represent independent components) of the various models were examined to identify what model resulted in the best solution. To determine the number of retainable components, the scree plot, which represents the number of substantive components, was examined (Kline, 2006). The communality (h^2), sum of the squared loadings per item, eigenvalues per component, or variance explained by each component were also estimated, where appropriate. Lastly, Cronbach's alpha indices were computed for each test component. Cronbach's alpha can be used to evaluate the internal consistency of test items, with a value greater than 0.70 indicating high consistency, 0.50 to 0.70 indicating moderate consistency, and a value below 0.50 indicating low consistency.

Confirmatory Factor Analysis (CFA)

AMOS (version 21) was used to perform a two-stage CFA on the second subset ($n = 281$) in order to test the fit of the optimal solution yielded in the PCA analysis. Initially independent CFA measurement models were generated for each component emerging in PCA. Each measurement model included a latent variable representing the sub-component measured by items alongside the items tapping that sub-component. Next, the complete CFA model was tested comprising the correlated sub-components and related items. Maximum Likelihood (ML) method of parameter estimation and multiple fit criteria to evaluate the fit of the model were employed as follows:

- (a) Chi-square test (χ^2): An index representing the difference between the observed and implied covariance or correlation matrices. Non-significant χ^2 values suggest good fit, although large samples can inflate this index.
- (b) Normed χ^2 (χ^2 / df): The ratio of χ^2 to the degrees of freedom (df). This ratio is small in well-fitting models (preferably below 3).
- (c) Two incremental indices: Non-Normed Fit Indices (NNFI) and Comparative Fit Indices (CFI). Both indices compare the postulated model to a baseline model that assumes that measures are not

correlated. Indices of 0.90 or above were chosen as indicators of satisfactory fit.

- (d) Root Mean Square Error of Approximation (RMSEA): This measure is used to adjust the significance level for chi-square tests when using large samples; a low RMSEA value is preferred.

To measure the effect of the constructs or dimensions on participants' performance, standardized regression estimates were used which quantify the increase in the standard deviation of the items if the standard deviation of the construct increases by one unit. Similarly, non-standard regression coefficients indicate the magnitude of increase in the items when the construct increases by one unit. The critical ratios (CRs) were estimated for each regression estimate by dividing them by their standard deviation. CRs greater than 1.96 are statistically significant ($p < 0.05$).

Rasch-Andrich Rating Scale Model (RSM)

RSM analysis (Andrich, 1978) were performed using the *WINSTEPS* computer package, Version 3.75 (Linacre, 2013a). Separate RSM analyses were performed on each subscale (component) and examined the hierarchy of item and person measures in each subscale. Reliability and separation estimations were examined, adherence to unidimensionality, and psychometric features of response categories in each subscale.

The fit of the data to the model was estimated by calculating the infit and outfit mean square (MNSQ) statistics. Infit MNSQ is sensitive to perturbations near the ability level of individuals or difficulty level (endorsability) of items. Outfit MNSQ is sensitive to outliers, meaning that if a participant with a low trait level endorses a high response category on a difficult item, she is an outlier inflating the outfit MNSQ (Brodersen, Thorsen, & Kreiner, 2007). Inliers are weighted so as to decrease this sensitivity and balance out the effect of outliers. MNSQ values should be equal to unity, though a narrow deviation from unity, that is 0.6 – 1.4, is acceptable (Bond & Fox, 2007). If MNSQ indices are lower than 0.6, the item does not provide much information and is redundant. Conversely, if the value is greater than 1.4, the item is inconsistent with the rest of the items and is likely contaminated by construct-irrelevant variance (Linacre, 2013b).

To investigate adherence to unidimensionality, a principal component analysis of Rasch linearized residuals (PCAR) was carried out. This analysis is

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

performed after the RSM dimension is extracted from the data and the researcher intends to examine potentially substantive structures in the residuals. If an identified structure has substance (eigenvalues > 2.5), it may represent a dimension beside the RSM dimension (Linacre, 2013b). Variance explained was assessed by the RSM dimension with the components which emerged in the PCAR.

The response category features, thresholds, and category measures of the affective dimension were examined. At threshold points, the adjacent response category curves intersect and participants with ability measures equal to or greater than the threshold begin to have a higher probability to select the higher category. Thresholds should increase monotonically from a lower to a higher index. Monotonic increase should ideally be larger than 1.4 but smaller than 5 logits (Bond & Fox, 2007).

Results

Descriptive Statistics of Data

Table 1 presents the descriptive statistics of the items including mean scores, standard deviation, skewness, and kurtosis. Items 29 and 19 had the lowest and highest mean values of 1.82 and 3.21, respectively. The skewness and kurtosis indices fall between -1 and +1, indicating that the sample size had a normal distribution.

Principal Component Analysis (PCA)

Multiple PCA on the first subset ($n = 282$) were performed. Due to space constraints, only the results of the optimal model in this section are reported. After conducting several PCA, Items 1, 4, 19, and 27 were deleted due to their cross-loading patterns in several models. The 26 items left loaded on three components which was called *affective* dimension, *behavioral* dimension, and *perseverance and confidence* dimension (correlation coefficients: .535, .329, and .440). To estimate this model, promax with Kaiser normalization rotation was used, which outperformed other rotation techniques. The Kaiser-Meyer-Olkin measure of sampling adequacy was 0.888, indicating that the variance in the data was caused primarily by the underlying construct and that PCA can be conducted.

Table 1. Descriptive Statistics of the Items ($n = 565$)

Item	Mean	Standard deviation	Skewness	Kurtosis
1	2.97	.869	-0.750	0.084
2	2.96	.738	-0.595	0.489
3	1.87	.683	0.506	0.387
4	2.94	.779	-0.665	0.401
5	2.81	.766	-0.380	-0.061
6	2.15	.795	0.218	-0.495
7	2.17	.766	0.130	-0.495
8	1.96	.683	0.579	0.835
9	2.81	.719	-0.410	0.202
10	2.73	.766	-0.327	-0.136
11	2.83	.743	-0.296	-0.114
12	2.08	.750	0.429	0.058
13	2.10	.745	0.402	0.038
14	2.05	.682	0.414	0.421
15	2.88	.735	-0.474	0.250
16	2.84	.702	-0.547	0.559
17	2.04	.787	0.478	-0.089
18	2.18	.791	0.278	-0.337
19	3.21	.763	-0.959	0.988
20	2.45	.737	0.328	-0.216
21	1.96	.717	0.526	0.360
22	2.20	.831	0.355	-0.366
23	2.03	.729	0.475	0.246
24	2.71	.755	-0.454	0.042
25	2.62	.783	-0.167	-0.355
26	2.66	.749	-0.270	-0.161
27	1.97	.740	0.555	0.295
28	2.71	.719	-0.224	-0.102
29	1.82	.774	0.942	0.930

This finding was further supported by the significant value of the Bartlett's test of sphericity which was 0.00 ($p < 0.05$), indicating that the correlation matrix was not an identity matrix—a matrix where all diagonal values are unity and off-diagonal indices are zero.

Table 2 presents the three-componential PCA model. Loading coefficients and communality values of all items (except Item 3) are greater than 0.300 indicating that the amount of variance explained by the components is significantly high. For example, Items 3, 6, 7, 8, 12, 13, 14, 17, 18, 21, 23, 26, and 29 loaded on the affective dimension with loading coefficients ranging from .458 to .884. Cronbach's alpha coefficients of the affective and behavioral dimensions are greater than .800, indicating significantly high internal consistency; perseverance and confidence dimension has a moderate Cronbach's alpha coefficient of .581.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 2. Rotated Factor Matrix Comprising Three Components Generated in PCA
(*n* = 282)

Item	Affective dimension	Behavioral dimension	Perseverance & confidence	Communality (<i>h</i> ²)
21	.886			.614
23	.757			.611
17	.749			.557
6	.629			.402
13	.608			.541
29	.599			.509
18	.579			.449
14	.568			.575
8	.563			.508
3	.534			.283
7	.472			.512
12	.459			.571
26	.566			.601
15		.777		.576
9		.749		.546
2		.746		.564
11		.744		.553
16		.506		.512
5		.482		.463
10		.481		.467
24			.737	.480
28			.627	.501
20			.512	.368
25			.452	.316
22			.378	.498
Cronbach's Alpha	.829	.858	.581	NA
Eigenvalues	8.080	1.896	1.397	NA
% Variance explained	32.32%	7.58%	5.59%	NA

*Note. Items 1, 4, 19, and 27 due to their cross-loading patterns in several models.

Confirmatory Factor Analysis (CFA) with AMOS

CFA was performed on the second subset (*n* = 283) to confirm the optimal PCA model across the second subset. The fit statistics of the measurement and full CFA models are presented in Table 3. The measurement models including the affective dimension, behavioral dimension, and perseverance and confidence dimension have good fit to the data. For example, the affective dimension's fit statistics are, as follows: $\chi^2 = 165.14$; $\chi^2/df = 2.54$; NNFI = 0.945; CFI = 0.960; and RMSEA = 0.052.

Table 3. Fit Statistics of Measurement and Three-Dimensional CFA Models ($n = 283$)

Model	χ^2	df	χ^2/df	NNFI	CFI	RMSEA	RMSEA boundaries
Affective dimension	165.14*	65	2.54	0.945	0.960	0.052	0.043–0.062
Behavioral dimension	13.612	14	.972	1.000	1.000	0.000	0.000–0.040
Perseverance & confidence dimension	306.65	15	.553	1.007	1.023	0.000	0.000–0.042
Unidimensional CFA model	994.11*	275	3.61	0.840	0.865	0.068	0.064–0.073
Three-dimensional CFA model	749.99*	272	2.75	0.912	0.910	0.056	0.051–0.061

*Note. * $p < 0.05$. $n = 281$.

Next, a unidimensional model was tested where all items were regressed on a general attitude dimension. The model had a poor fit to the data ($\chi^2 = 994.11$; $\chi^2/df = 3.61$; NNFI = 0.840; CFI = 0.865; and RMSEA = 0.068) and accordingly the fit of a three-dimensional CFA model composed of separate affective, behavioral, and perseverance and confidence dimensions was assessed, which fitted the data well ($\chi^2 = 749.99$; $\chi^2/df = 2.75$; NNFI = 0.912; CFI = 0.910; and RMSEA = 0.056). The χ^2 value of this model is significant at $p < 0.05$, which can be attributed to the sample size as well the complexity of the model as indicated by the degrees of freedom ($df = 272$).

Table 4 displays standard and non-standard regression estimates, standard error of measurement, CRs, and their p -values of the three-dimensional CFA model. All items have significantly high regression estimates as testified by their CRs and p -values. For example, Item 3 measures the affective dimension with a non-standard regression estimate of 0.750 ($p < 0.001$; CR = 9.738), suggesting that if participants' affective dimension increases by one unit, their performance on that item will increase by 0.75 units.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 4. Item Statistics of the Three-Dimensional CFA ($n = 283$)

Item	Dimension	Non-standardized estimate	Standardized estimate	SE	CR	p-value
3	Affective	.750	.470	.077	9.738	***
6	Affective	.986	.531	.091	10.797	***
7	Affective	1.067	.597	.090	11.798	***
8	Affective	.937	.587	.080	11.670	***
12	Affective	1.257	.719	.093	13.529	***
13	Affective	1.219	.699	.092	13.284	***
14	Affective	1.178	.742	.085	13.817	***
17	Affective	1.225	.667	.095	12.830	***
18	Affective	1.086	.587	.093	11.653	***
21	Affective	1.000	.597	Constrained for parameter estimation		
23	Affective	1.176	.691	.089	13.147	***
26	Affective	1.018	.582	.088	11.593	***
29	Affective	1.203	.665	.094	12.826	***
2	Behavioral	.956	.667	.066	14.495	***
5	Behavioral	.972	.654	.069	14.167	***
9	Behavioral	.920	.660	.064	14.307	***
10	Behavioral	.989	.666	.069	14.404	***
11	Behavioral	.987	.684	.067	14.822	***
15	Behavioral	1.000	.699	Constrained for parameter estimation		
16	Behavioral	.942	.691	.063	14.961	***
20	PersConf	1.287	.463	.192	6.702	***
22	PersConf	2.143	.684	.280	7.646	***
24	PersConf	1.000	.351	Constrained for parameter estimation		
25	PersConf	1.380	.468	.205	6.726	***
28	PersConf	1.488	.550	.208	7.160	***

*Note. PersConf = Perseverance & confidence dimension. CR = Critical ratio. SE = standard error of measurement. *** $p < 00.1$.

Finally, [Figure 1](#) displays the three-dimensional CFA model with correlation coefficients (bidirectional arrows connecting the larger circles on the left side of the model) between the three dimensions, which indicate they are highly related. The figure also displays standardized regression coefficients which are just above the one headed arrows moving from dimensions (large circles) to items, represented by rectangles. Errors of measurement are represented as small circles on the right side of the figure.

Figure 1. Graphic representation of the three-dimensional CFA model.
(PersConf = perseverance and confidence dimension)

Rasch-Andrich Rating Scale Model (RSM)

Affective Dimension

The RSM was initially applied to the affective dimension consisting of 13 items. Person ability (or trait levels) ranged between -4.27 and 2.34 logits ($M = -0.96$; $SD = 1.24$) and item endorsability measures ranged between -1.46 and 0.74 logits ($M = 0.00$; $SD = 0.52$). [Figure 2](#) presents an item person map with Andrich thresholds, which is a more precise map for polytomous data than a map without the thresholds. The left column represents respondents and the right column demonstrates the items located at their Andrich thresholds—points above which respondents begin to have a higher probability to choose higher categories. For example, Item 26's (x26) first threshold represented as 0.2 is the lowest of all and any respondent who has a higher ability (affective dimension measure) would have a higher probability to choose scoring category 3 on the item, whereas respondents below that point would most likely choose category 1. There is a gap between 0 and -1 logits, although there are not many respondents with trait levels corresponding to it and there are sufficient thresholds distinguishing respondents on the map.

Rasch model item and person reliability indices for this dimension were .98 (separation = 6.86) and .80 (separation = 1.94), respectively, indicating that if the items are administered to another group with similar features from the same population, there is high confidence that item and respondent measures (or locations) are reproducible ([Linacre, 2013b](#)).

[Table 5](#) presents fit statistics, item endorsability measures, and raw scores for affective dimension items. The most lowly endorsed item (most difficult to endorse) is Item 13 (measure = 0.74), and most highly endorsed item is Item 12 (measure = -1.46). Infit MNSQ values range between 0.749 (Item 23) and 1.073 (Item 6) and outfit MNSQ values range between 0.719 (Item 23) and 1.080 (Item 6). This provides evidence that the data has not been contaminated by construct-irrelevant variance and high and low performers responded to the items as expected by the RSM.

Figure 2. The affective dimension's item person map with Andrich thresholds. The map is truncated at the bottom.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 5. Raw Scores, Fit Statistics, and Item Endorsability Measures of Affective Dimension Items

Item	Rasch model measure	Raw score	Infit MNSQ	Outfit MNSQ
3	0.59	1043	1.012	1.026
6	-0.19	1204	1.073	1.080
7	-0.25	1206	0.897	0.905
8	0.31	1099	0.787	0.766
12	-0.01	1158	0.971	0.951
13	-0.07	1174	0.765	0.735
14	0.07	1137	0.978	0.861
17	0.1	1138	0.895	0.858
18	-0.27	1209	1.003	1.019
21	0.33	1094	0.834	0.786
23	0.11	1129	0.749	0.719
26	-1.46	1492	1.041	1.078
29	0.74	1021	1.032	0.965

Next, PCAR was performed to further explore the dimensionality of the test. The analysis showed that the raw variance explained by items and persons was 38.2% (7.4 eigenvalue units), whereas the first dimension in residuals explained 8.6% (1.8 eigenvalue units), indicating that the structure in residuals has no substance, as it accounts for a tiny proportion of the observed variance.

Finally, the response category features, thresholds, and category measures of the affective dimension were examined. Thresholds in this dimension increase monotonically (threshold 1 = -2.50; threshold 2 = 0.30; threshold 3 = 2.20). In addition, each response category has at least 10 observed counts and acceptable fit MNSQ indices (infit MNSQ: category 1 = 0.89; category 2 = 0.90; category 3 = 0.85; category 4 = 1.45), indicating that participants selected categories as expected by the model.

Behavioral Dimension

The item and person reliability indices in this dimension were .99 (separation = 1.47) and .69 (separation = 10.26), respectively. The item person map with Andrich thresholds is also displayed in [Figure 3](#). This figure shows a noticeable gap around the mean score. Otherwise, the spread of item thresholds corresponds to respondents' trait levels, though adding one or two items to fill in the gap would enhance the reliability of the dimension.

Figure 3. The behavioral dimension's item person map with Andrich thresholds. The map is truncated at the bottom.

Table 6 gives fit statistics, item endorsability measures, and raw scores for behavioral dimension items. Item measures ($M = 0.00$; $SD = 0.80$) range between -0.67 (Item 2) and 2.06 (Item 8) and person measures ($M = 0.09$; $SD = 0.63$) range between -5.04 and 3.27. Infit and outfit MNSQ statistics range between 0.70 and 1.35, suggesting that the items all fit the model, and that the data has not been contaminated by perturbations.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 6. Raw Scores, Fit Statistics, and Item Endorsability Measures of Behavioral Dimension Items

Item	Rasch model measure	Raw score	Infit MNSQ	Outfit MNSQ
2	-0.67	1663	0.864	0.827
5	-0.22	1561	0.904	0.895
9	-0.22	1566	0.954	0.930
8	2.06	1099	1.350	1.242
10	0.03	1511	0.816	0.999
11	-0.26	1580	0.982	0.964
15	-0.42	1609	0.884	0.856
16	-0.28	1585	0.717	0.687

Subsequently, PCAR was performed on the Rasch model residuals of this dimension to examine the presence of any substantive structure. The Rasch model item and person measures explained 44.5% of the observed variance, extracting 6.2 eigenvalues. The first contrast extracted only 1.5 eigenvalues, explaining 18.9% of the variance in data which is not a substantive amount (Linacre, 2013b). These results provide further evidence supporting the unidimensionality of this dimension.

Finally, the response category functions were examined. It was found a monotonic increment of the thresholds at -2.54, -0.23, and +2.77, with proper category infit MNSQ indices (1.25, 0.68, 0.91, 0.99). The fit statistics and ascending category measures provide evidence supporting the measurement features of response categories.

Perseverance and Confidence Dimension

Respective item and person reliability indices are 0.98 (separation = 6.25) and 0.41 (separation = 1.31). The rather low person reliability is likely due to the small number of items.

Figure 4 presents perseverance and confidence dimension's item person map with Andrich thresholds. There are two rather small gaps in the map, although there are not many respondents with trait levels corresponding to the gaps and there are sufficient thresholds distinguishing respondents.

Figure 4. The perseverance and confidence dimension's item person map with Andrich thresholds.

Table 7 fit statistics, item endorsability measures, and raw scores for perseverance and confidence dimension items. The most difficult item to endorse is Item 22 (measure = 0.75) and the easiest is Item 28 (measure = -0.39). Infit and outfit MNSQ statistics fall between 0.7 to 1.20, suggesting that the items were not contaminated by construct-irrelevant factors.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 7. Raw Scores, Fit Statistics, and Item Endorsability Measures of Perseverance and Confidence Dimension Items

Item	Rasch model measure	Raw score	Infit MNSQ	Outfit MNSQ
20	0.20	1361	0.768	0.773
22	0.75	1228	1.106	1.151
24	-0.37	1500	0.819	0.817
25	-0.18	1453	0.890	0.904
28	-0.39	1506	0.711	0.713

Next, the dimension's residuals were examined through PCAR. Item and person measures explain 27% of the variance or 4.4 eigenvalue units, whereas 14.7% of the variance in the data or 1.00 eigenvalue units is accounted for by the first contrast. This further supports the unidimensionality of the perseverance and confidence dimension.

Finally, the functionality of response categories was examined. Like the other two dimensions, the thresholds had a monotonic increase: -2.03, -0.08, and 2.10. Infit and outfit MNSQ of the response categories fell between 0.80 and 1.20. This analysis provides support for the functionality of response categories.

Discussion

The present study was designed to develop and build a validity argument for P-BAQ. In this section, the findings and their contribution to the theory of attitude are discussed and are subsequently developed into a validity argument for the instrument.

P-BAQ's Dimensions

The data analysis procedures provided evidence supporting the presence of three dimensions underlying P-BAQ: affective, behavioral, and perseverance and confidence dimensions. Consistent with previous research, the present study yields support for the multidimensionality of attitude (e.g., [Eagly & Chaiken, 1993](#); [Hogg & Vaughan, 2011](#); [McLeod, 2009](#); [Mantle-Bromley, 1995](#)). The results also lend partial support to the attitude's basic model, which comprises affective, behavioral, and cognitive dimensions ([McLeod, 2009](#)).

However, unlike the basic model, the study did not yield a cognitive dimension. This is in line with [Kay \(1993\)](#), [Mantle-Bromley \(1995\)](#), [Min \(1998\)](#),

Selwyn (1997), and Tsai et al. (2001) who conceptualize attitude as a set of “evaluative” responses to individuals, objects, or opinions but distinguish it from a cognitive dimension measuring individuals’ beliefs and perceptions—although cognitive capacity is a precondition to forming and holding attitudes. Relatedly, Ajzen (2005, 2011) argues that there is considerable commonality between cognitive and affective dimensions, and that is why distinguishing the two components might be impossible. The partial disagreement over the constituent structure of attitude may also be attributed to the lack of clarity surrounding attitude’s definition, as there is still no universally endorsed definition for attitude (Franzoi, 2003; Smith et al., 2000).

Findings further show that attitude toward blogs is highly similar to attitude toward computers, internet, and web-based technologies such as social networking sites (e.g., Fančovičová & Prokop, 2008; Liaw, 2002; Sam et al., 2008). However, the emergence of the perseverance and confidence dimension is new in this study. If attitude is conceptualized as a continuum with affective and behavioral ends, the perseverance and confidence dimension would fall between the two ends. It might be said that the individuals who score low on this dimension would likely score low on both affective and behavioral dimensions and vice versa; that is, performance on this dimension might predict students’ performance on the other two dimensions. Future research can examine this possibility by using causal statistical models.

Our findings undermine Daud’s (1995) and Smith’s (1996) claims of attitude’s unidimensionality. The results of PCA showed three separate dimensions, and CFA model testing confirmed that a unidimensional model would not fit the data. Although the correlation coefficients between the three dimensions were high in the CFA model, high correlations may not be confused with identity of dimensions (Borsboom, 2008); two independent concepts such as age and height might be highly correlated yet distinct in definition and nature. It should be noted that dimensions are *not* psychological traits but a rank-ordering of people on the traits. Aryadoust (2013, p. 195) stated:

Using the example of two “climbing” [dimensions], “holding onto ledges” and “pulling oneself up from ledges,” the correlation between two factors represents how well individual differences on one sub-skill (holding onto ledges) *linearly* predicts individual differences on another (pulling oneself up from ledges). Whether that correlation is 0, 0.5, or 1 is immaterial to the question of whether or not the [dimensions] exist. It only addresses the

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

question whether individual differences on one predict individual differences on another.

Despite the high psychometric quality of the three dimensions, it would be useful to add a few items to each dimension to improve person reliability statistics. Overall, the multidimensional structure of attitude toward blogs is supported in the present study. A validity argument for P-BAQ will be developed in the following section using the evidence yielded at different stages of the study.

Validity Argument of P-BAQ

A validity argument provides a coherent treatment of the evidence gathered in the validation study. P-BAQ's validity argument is supported by the findings of the previously stated psychometric studies and is presented in [Table 8](#). For example, the domain inference is based on the claim concerning the multidimensionality of P-BAQ. The claim is based on the assumptions that dimensions of P-BAQ can be determined and measured. It is supported by the postulations that "Attitude toward blogs is composed of multiple components which were identified through the survey of literature in this study" and "The pool of items adapted from previous attitude research contained a wide range of items tapping attitude's dimensions." The presented literature survey shows that the commonly adhered-to attitude construct is multidimensional (e.g., [Selwyn, 1997](#); [Tsai et al., 2001](#)) and supports the representativeness of the item pool. In addition, items chosen for translation by experts would engage at least two dimensions of attitudes: affective and behavioral. However, [Table 8](#) presents a potential counterargument: some researchers provided empirical evidence supporting attitude's unidimensionality. The counterargument is rebutted by the evidence from PCA and CFA, thereby supporting the domain definition inference and the theoretical representation of P-BAQ.

The remainder of the inferences are also supported by the available evidence. However, the generalization inference is undermined by low person reliability and separation statistics of the three dimensions which result from the gaps in the item person map. This would indicate a requirement to add a few items to P-BAQ to enhance the generalizability of individuals' attitude scores.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Table 8. Validity argument of the P-BAQ

Inference	Claim	Assumption	Postulations (warrants)	Backing
Domain definition	Iranian tertiary students' attitude toward blogs is a multidimensional construct including at least two dimensions: affective and behavioral.	(i) The components (dimensions) of attitude toward blogs can be determined; and (ii) the components are measurable.	(a) Attitude toward blogs is composed of multiple components which were identified through the survey of literature in this study; and (b) the pool of items adapted from previous attitude research contained a wide range of items tapping attitude's dimensions.	The presented literature survey shed light on the theory of attitude toward technology and supports the thoroughness of the item pool created. The literature also supported multidimensionality of attitude although a handful of studies have yielded unidimensional attitude scales. Finally, items which are chosen for translation by experts would engage at least two dimensions of attitudes: affective and behavioral.
Translation	The chosen items are clearly translated into Persian.	(i) The items are translatable; and (ii) the translated items are easy to understand and contain no ambiguous terms or phrases for Persian speakers.	Translation would have no adverse effect on the construct equivalence of the Persian instrument if it is approved by experts.	The approval of translation specialists of the clarity of the translated items and their lack of linguistic ambiguities was obtained.
Evaluation	Participants consistently choose their responses, producing reliable data.	Scoring categories (Likert scale) are appropriate and precise and thus can be established through psychometric and statistical analysis.	(a) Four-category Likert scales would prevent respondents from "flat-lining" the scale which happens when they endorse the mid-points on the scale; and (b) Psychometric attributes of the Likert scale are backed by psychometric analysis.	(1) Monotonic increment of Likert scales' response categories and thresholds was established; and (2) appropriate fit statistics for each response category were supported by RSM.

Table 8 continued on the next page

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Generalization	The observed scores can be generalized to the universe of scores.	The performance on the survey is considered representative of performance on the universe of items.	Generalization is psychometrically measured.	High Rasch measurement scores on the item reliability and separation indices support that the observed measures can be generalized to the universe of measures. <u>Undermining evidence:</u> In contrast, low person reliability and separation indices on the three dimensions suggest that additional items should be added to the P-BAQ.
Explanation	A link is established between the observed scores and the underlying construct.	(i) Construct representativeness of P-BAQ can be examined; (ii) the correlation of the dimensions can be examined; and (iii) lack of construct-irrelevant factors can be investigated.	(a) The constituent structure of P-BAQ is supported by theory; (b) the sub-components (dimensions) of P-BAQ are distinguishable and related; and (c) P-BAQ has reliable psychometric features and dimensionality analysis can show that its components are not contaminated.	The results of psychometric analysis: PCA supported the multidimensional structure of the instrument; CFA verified the structure across a different sample, supporting the psychometric distinction of the sub-components; and RSM results lent themselves to the unidimensionality of the dimensions of P-BAQ, disproving the presence of construct-irrelevant factors. However, the item person map in the RSM analysis contained a few gaps, which should be filled in by further items in the future.

Applications of P-BAQ

Turning to the pedagogical overtones of the study, research shows that learners' progress largely depends on their motivation, which is a "composite of attitudinal variables" (Mantle-Bromley & Miller, 1991, p. 418). Positive attitudes toward learning tools such as blogs and classes can improve students' motivation. Measuring students' attitude toward blogs will give an indication of the potential causes of students' lack of motivation and help teachers take proper measures to motivate the students.

To our knowledge, P-BAQ is the first Persian blog attitude instrument that has a (strong) validity argument. Because it is a valid psychometric tool, the P-BAQ is suitable for use in educational programs that make use of blogs. It is also of potential use for research conducted in Afghanistan, Iran, Tajikistan, and other countries where Persian or Farsi are spoken.

Conducting continuous research into the attitude of learners who use blogs in low-resource countries is of paramount importance. It might be said that a lack of facilities and financial aid coupled with cultural factors and unfamiliarity with blogs (and other technology advances) can affect students' attitudes and motivations, and consequently their learning achievements.

Conclusion

A validity argument for the P-BAQ was constructed by examining claims and assumptions, proposing postulations, and offering backing (evidence) supporting these postulations. Overall, the validity argument is strong, though the structure of the instrument could be improved through future research. Validation of psychometric instruments is never completed; it is extended and "refined as new research and data analysis technologies are developed" (Aryadoust, 2013, p. 223). Future research can address other issues such as bias or differential item functioning which are related to the explanation inference.

Articulating the claims concerning the psychometric quality of instruments as well as uses and interpretations of scores would benefit validation research. However, research into measuring attitude has hardly sought to construct and assess validity arguments. To our knowledge, the only study attempting to construct a validity argument for a survey instrument was done by Bangert (2009). It is hoped that validity arguments will be applied more often in future attitude research and other educational studies.

References

- Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561-573. doi:[10.1007/BF02293814](https://doi.org/10.1007/BF02293814)
- Al-Tamimi, A., & Shuib, M. (2009). Motivation and attitudes towards learning English: A study of petroleum engineering undergraduates at Hadhramout University of Sciences and Technology. *GEMA Online Journal of Language Studies*, 9(2), 29-53.
- Ajzen, I. (2005). *Attitudes, personality, and behaviour* (2nd ed.). New York: Open University Press.
- Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. *Psychology & Health*, 26(9), 1113-1127. doi:[10.1080/08870446.2011.613995](https://doi.org/10.1080/08870446.2011.613995)
- Aryadoust, V. (2013). *Building a validity argument for a listening test of academic proficiency*. Newcastle: Cambridge Scholars Publishing.
- Aryadoust, V., Mehran, P., & Alizadeh, M. (2014). Validating a computer-assisted language learning attitude instrument used in Iranian EFL context: An evidence-based approach. *Computer Assisted Language Learning Journal-First Online*. doi:[10.1080/09588221.2014.1000931](https://doi.org/10.1080/09588221.2014.1000931)
- Ayres, R. (2002). Learner attitudes towards the use of CALL. *Computer Assisted Language Learning*, 15(3), 241-249. doi:[10.1076/call.15.3.241.8189](https://doi.org/10.1076/call.15.3.241.8189)
- Bangert, A. W. (2009). Building a validity argument for the community of inquiry survey instrument. *Internet and Higher Education*, 12(2), 104-111. doi:[10.1016/j.iheduc.2009.06.001](https://doi.org/10.1016/j.iheduc.2009.06.001)
- Blackstone, B., Spiri, J., & Naganuma, N. (2007). Blogs in English language teaching and learning: Pedagogical uses and student responses. *Reflections on English Language Teaching*, 6(2), 1-20.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates.
- Borsboom, D. (2008). Latent variable theory. *Measurement: Interdisciplinary Research and Perspectives*, 6(1/2), 25-53. doi:[10.1080/15366360802035497](https://doi.org/10.1080/15366360802035497)
- Brodersen, J., Thorsen, H., & Kreiner, S. (2007). Validation of a condition-specific measure for women having an abnormal screening mammography. *Value in Health*, 10(4), 294-304. doi:[10.1111/j.1524-4733.2007.00184.x](https://doi.org/10.1111/j.1524-4733.2007.00184.x)
- Chen, W. C., Shih, Y., & Liu, G. Z. (2013). Learning task design and its induced learning effects in a cross-institutional blog-mediated telecollaboration.

Computer Assisted Language Learning, 28(4), 285-305.

doi:[10.1080/09588221.2013.818557](https://doi.org/10.1080/09588221.2013.818557)

Chu, S.-C., & Kamal, S. (2008). The effect of perceived blogger credibility and argument quality on message elaboration and brand attitudes: An exploratory study. *Journal Interactive Advertising*, 8(2), 32-33.

doi:[10.1080/15252019.2008.10722140](https://doi.org/10.1080/15252019.2008.10722140)

Daud, N. M. (1995). A computer attitude scale for language teachers. *Computer Assisted Language Learning*, 8(4), 355-363.

doi:[10.1080/0958822950080404](https://doi.org/10.1080/0958822950080404)

DeVellis, R. F. (2003). *Scale development: Theory and applications* (2nd ed.). Thousand Oaks, London, New Delhi: Sage Publications.

Dimroth, C. (2010). The acquisition of negation. In L. R. Horn (Ed.), *The expression of negation* (pp. 39-73). Berlin/New York: Mouton de Gruyter.

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing* (2nd ed.). New York: Routledge.

Duda, G., & Garrett, K. (2008). Blogging in the physics classroom: A research-based approach to shaping students' attitudes toward physics. *American Journal of Physics*, 76(11), 1054-1065. doi:[10.1119/1.2967707](https://doi.org/10.1119/1.2967707)

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes*. Fort Worth, TX: Harcourt Brace Jovanovich.

Fageeh, A. I. (2011). EFL learners' use of blogging for developing writing skills and enhancing attitudes towards English Learning: An exploratory study, *Journal of Language and Literature*, 2(1), 2078-0303.

Fančovičová, J., & Prokop, P. (2008). Students' attitudes toward computer use in Slovakia. *Eurasia Journal of Mathematics, Science & Technology Education*, 4(3), 255-262.

Farahzad, F. (1992). Testing achievement in translation classes. In C. Dollerup & A. Loddergard (Eds.), *Teaching translation and interpreting: Training, talent and experience* (pp. 271-278). Amsterdam/Philadelphia: John Benjamins.

Franzoi, S. (2003). *Social Psychology*. Boston: McGraw-Hill.

Gajewski, J. R. (2007). Neg-raising and polarity. *Linguistics and Philosophy*, 30(3), 289-328. doi:[10.1007/s10988-007-9020-z](https://doi.org/10.1007/s10988-007-9020-z)

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). New Jersey: Pearson Educational Product.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

- Harkness, J. A. (2008). Comparative survey research: Goals and challenges. In E. D. De Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 56-77). New York, NY: Lawrence Erlbaum.
- Harwood, C. (2010). Using blogs to practice grammar editing skills. *English Language Teaching World Online*, 2. Retrieved from <http://blog.nus.edu.sg/eltwo/2010/08/11/using-blogs-to-practice-grammar-editing-skills/#more-1107>
- Hogg, M. A., & Vaughan, G. M. (2011). *Social psychology* (6th ed.). Harlow: Pearson.
- Huang, H. C. (2013). From web-based readers to voice bloggers: EFL learners' perspectives. *Computer Assisted Language Learning*, 28(2), 145-170. doi:10.1080/09588221.2013.803983
- IBM Corp. (2012). IBM SPSS Statistics for Windows, Version 21.0 [Computer Software]. Armonk, NY: IBM Corp.
- Jones, T., & Clarke, V. A. (1994). A computer attitude scale for secondary students. *Computers & Education*, 22(4), 315-318. doi:10.1016/0360-1315(94)90053-1
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Westport: American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. doi:10.1111/jedm.12000
- Kay, R. H. (1993). An exploration of theoretical and practical foundations for assessing attitudes towards computers: The Computer Attitude Measure (CAM). *Computers in Human Behavior*, 9(4), 371-386. doi:10.1016/0747-5632(93)90029-R
- Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.
- Kline, R. (2006). Reverse arrow dynamics: Formative measurement and feedback loops. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 43-68). Greenwich, CT: Information Age Publishing.
- Liaw, S.-S. (2002). An Internet survey for perceptions of computer and the World Wide Web: Relationship, prediction, and difference. *Computers in Human Behaviour*, 18(1), 17-35. doi:10.1016/S0747-5632(01)00032-2
- Liaw S.-S., Huang, H. M., Chen, G. D. (2007). Surveying instructor and learner attitudes toward e-learning. *Computers & Education*, 49(4), 1066-1080.

- Linacre, J. M. (2013a). WINSTEPS Rasch measurement [computer program]. Chicago: Winsteps.com.
- Linacre, J. M. (2013b). *A user's guide to WINSTEPS*. Chicago: Winsteps.com.
- Loyd, B. H., & Gressard, C. P. (1984). Reliability and factorial validity of computer attitude scale. *Educational and Psychological Measurement*, 44(2), 501-505. doi:10.1177/0013164484442033
- Mantle-Bromley, C. (1995). Positive attitudes and realistic beliefs: Links to proficiency. *The Modern Language Journal*, 79(3), 371-386. doi:10.1111/j.1540-4781.1995.tb01114.x
- Mantle-Bromley, C., & Miller, R. B. (1991). Effect of multicultural lessons on attitudes of students of Spanish. *The Modern Language Journal*, 75(4), 418-425. doi:10.1111/j.1540-4781.1991.tb05376.x
- McIntosh, E. (2009). Blogging improves young people's confidence in their writing and reading. Retrieved from <http://edu.blogs.com/edublogs/2009/12/blogging-improves-young-peoples-confidence-writing-reading.html>
- McLeod, S. A. (2009). *Attitudes and behaviour*. Retrieved from <http://www.simplypsychology.org/attitudes.html>
- Min, B. C. (1998). *A study of the attitudes of Korean adults toward technology assisted instruction in English language programs* (Unpublished doctoral dissertation). University of Northern Illinois, DeKalb.
- Morse, B. J., Gullekson, N. L., Morris, S. A., & Popovich, P. M. (2011). The development of a general Internet attitudes scale. *Computers in Human Behavior*, 27(1), 480-489. doi:10.1016/j.chb.2010.09.016
- Pektas, S. T., & Erkip, F. (2006). Attitudes of Design Students toward Computer Usage in Design. *International Journal of Technology and Design Education*, 16(1), 79-95. doi:10.1007/s10798-005-3175-0
- Pickens, J. (2005). *Attitudes and perceptions*. Retrieved from <http://healthadmin.jbpub.com/borkowski/chapter3.pdf>
- Sam, H. K., Othman, A. E. A., & Nordin, Z. S. (2005). Computer self-efficacy, computer anxiety, and attitudes toward the Internet: A study among undergraduates in Unimas. *Journal of Educational Technology & Society*, 8(4), 205-219.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

- Selwyn, N. (1997). Students' attitudes toward computers: Validation of a computer attitude scale for 16–19 education. *Computers & Education*, 28(1), 35-41. doi:[10.1016/S0360-1315\(96\)00035-8](https://doi.org/10.1016/S0360-1315(96)00035-8)
- Shaft, T. M., Sharfman, M. P., & Wu, W. W. (2004). Reliability assessment of the attitude towards computers instrument (ATCI). *Computers in Human Behaviour*, 20(5), 661-689. doi:[10.1016/j.chb.2003.10.021](https://doi.org/10.1016/j.chb.2003.10.021)
- Shahsavar, Z., & Tan, B. H. (2011). Does cognitive style affect bloggers' attitude in an online learning environment? *GEMA Online TM Journal of Language Studies*, 11(1), 159-171.
- Shahsavar, Z., & Tan, B. H. (2012). Developing a questionnaire to measure students' attitudes toward the course blog. *Turkish Online Journal of Distance Education*, 13(1), 200-210.
- Smith, R. M. (1996). A comparison of methods for determining dimensionality in Rasch measurement. *Structural Equation Modelling*, 3(1), 25-40. doi:[10.1080/10705519609540027](https://doi.org/10.1080/10705519609540027)
- Smith, B., Caputi, P., & Rawstorne, P. (2000). Differentiating computer experience and attitudes toward computers: An empirical investigation. *Computers in Human Behavior*, 16(1), 59-81. doi:[10.1016/S0747-5632\(99\)00052-7](https://doi.org/10.1016/S0747-5632(99)00052-7)
- Teo, T. (2012). Measurement invariance of self-reported data across gender: A study of technology acceptance. *International Journal of Educational and Psychological Assessment*, 11(1), 43-54.
- Tsai, C.-C., Lin, S. S. J., & Tsai, M.-J. (2001). Developing an Internet attitude scale for high school students. *Computers & Education*, 37(1), 41-51. doi:[10.1016/S0360-1315\(01\)00033-1](https://doi.org/10.1016/S0360-1315(01)00033-1)
- Tu, C.-C., Chen, P.-C., & Lee, M.-Y. (2007). Fostering EFL learners' writing competence through web-Based guided writing. *WHAMPOA—An Interdisciplinary Journal*, 53, 225-244.
- Vandewaetere, M., & Desmet, P. (2009). Introducing psychometrical validation of questionnaires in CALL research: The case of measuring attitude towards CALL. *Computer Assisted Language Learning*, 22(4), 349-380. doi:[10.1080/09588220903186547](https://doi.org/10.1080/09588220903186547)
- Wang, Q., & Woo, H. L. (2010). Investigating students' critical thinking in weblogs: An exploratory study in a Singapore secondary school. *Asia Pacific Education Review*, 11(4), 541-551. doi:[10.1007/s12564-010-9101-5](https://doi.org/10.1007/s12564-010-9101-5)

- Woo, H. L., & Wang, Q. (2009). Using weblog to promote critical thinking: An exploratory study. *World Academy of Science, Engineering and Technology, International Science Index 25, International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering*, 3(1), 48-57.
- Wright, B. D. (1994a). Comparing factor analysis and Rasch measurement. *Rasch Measurement Transactions*, 8(1), 350.
- Wright, B. D. (1994b). Local dependency, correlations, and principal components. *Rasch Measurement Transactions*, 10(3), 509-511.
- Zawilinski, L. (2009). HOT blogging: A framework for blogging to promote higher order thinking. *The Reading Teacher*, 62(8), 650-661.
doi:[10.1598/RT.62.8.3](https://doi.org/10.1598/RT.62.8.3)
- Zhang, Y. (2007). Development and validation of an Internet use attitude scale. *Computers & Education*, 49(2), 243-253.
doi:[10.1016/j.compedu.2005.05.005](https://doi.org/10.1016/j.compedu.2005.05.005)
- Zhao, Y., & Frank, K. A. (2003). Factors affecting technology uses in school: An ecological perspective. *American Educational Research Journal*, 40(4), 807-840. doi:[10.3102/00028312040004807](https://doi.org/10.3102/00028312040004807)
- Zimbardo, P. G., & Leippe, M. R. (1991). *The psychology of attitude change and social influence*. New York: McGraw-Hill.

VALIDATING BLOG ATTITUDE QUESTIONNAIRE

Appendix

P-BAQ's items

	کاملاً مخالف	مخالف	موافق	کاملاً موافق
۱				کار کردن با وبلاگ مرا مضطرب نمی کند.
۲				کار کردن با وبلاگ را دوست دارم.
۳				کار کردن با وبلاگ مرا عصبی می کند.
۴				وقتی کسی در مورد وبلاگ صحبتی به میان می آورد احساس نگرانی نمی کنم.
۵				انتخاب درسهایی که در رابطه با وبلاگ است، برای من خوشایند هست.
۶				با وبلاگ نمی توانم خوب کار کنم.
۷				تبادل اطلاعات از طریق وبلاگ برایم خوشایند نیست.
۸				از کار کردن با وبلاگ احساس ناراحتی می کنم.
۹				معمولاً وقتی برای تبادل اطلاعات از وبلاگ استفاده می کنم، احساس رضایت دارم.
۱۰				در کلاسهایی که از وبلاگ استفاده می شود، احساس راحتی می کنم.
۱۱				فکر می کنم کار کردن با وبلاگ خیلی لذت بخش و جالب باشد.
۱۲				فکر نمی کنم از کار کردن با وبلاگ لذت ببرم.
۱۳				تبادل نظر با دیگران از طریق وبلاگ برای من خوشایند نیست.
۱۴				وقتی به کار کردن با وبلاگ فکر می کنم، احساس خوبی ندارم.
۱۵				مطمئن هستم که وبلاگ در انجام دادن کارهایم مفید خواهد بود.
۱۶				از کار کردن با وبلاگ احساس رضایت میکنم.
۱۷				من برای کار کردن با وبلاگ ساخته نشده ام.
۱۸				برای من قابل توجه نیست که چگونه بعضی از افراد زمان زیادی را صرف کار کردن با وبلاگ میکنند و از کار کردن با آن لذت می برند.
۱۹				مطمئنم که می توانم کار کردن با وبلاگ را یاد بگیرم.
۲۰				وقتی که شروع به کار کردن با وبلاگ می کنم، دوست ندارم از آن دست بکشم.
۲۱				نصرت می کنم که کار کردن با وبلاگ برای من دشوار باشد.
۲۲				در مجموع تمایل چندانی به استفاده از وبلاگ ندارم.
۲۳				از کار کردن با وبلاگ احساس سردرگمی و نارضایتی می کنم.
۲۴				اگر مسئله ای در حین استفاده از وبلاگ در کلاس حل نشده باقی بماند، بعد از کلاس به آن فکر میکنم.
۲۵				در درسهایی که مربوط به وبلاگ است، می توانم نمرات خوبی کسب کنم.
۲۶				از صحبت با دیگران در مورد وبلاگ لذت می برم.
۲۷				گذراندن درسهای مربوط به وبلاگ از عهده من خارج است.
۲۸				زمان کار کردن با وبلاگ از اعتماد به نفس بالایی برخوردارم.
۲۹				نسبت به وبلاگ احساس انزجار و تنفر می کنم.

The English Version of the Blog Attitude Questionnaire

#	Affective	SD	D	A	SA
3	Keeping a blog would make me very nervous.				
6	I'm no good with blogs.				
7	The challenge of adding a new post on blogs and keeping them updated does not appeal to me.				
8	Blogs make me feel uncomfortable.				
12	I don't think I would enjoy doing advanced blog work (such as designing complex appearance).				
13	Figuring out any blog problem does not appeal to me.				
14	I get a sinking feeling when I think of trying to keep a blog.				
17	I'm not the type to do well with blogs.				
18	I don't understand how some people can spend so much time working with blogs and seem to enjoy it.				
21	I think using or keeping a blog would be very hard for me.				
23	Keeping blogs make me feel uneasy and confused.				
26	I don't enjoy talking with others about blogs.				
29	I feel aggressive and hostile toward blogs.				
	Behavioral				
2	I would like working with and keeping blogs.				
5	It wouldn't bother me at all to take courses on blogs.				
9	Generally, I would feel OK about adding a new post on the blog.				
10	I would feel at ease in a blog class.				
11	I think working with and keeping blogs would be enjoyable and stimulating.				
15	I am sure I could do work with blogs.				
16	I would feel comfortable working with a blog (e.g., keeping and updating them, referencing to other sources, etc.).				
	Perseverance and Confidence				
20	Once I start to work on a blog, I would find it hard to stop.				
22	I will do as little work through blogs as possible.				
24	If a problem with my blog is left unsolved, I would continue to think about it afterward.				
25	I could get good grades in blog courses if there are any.				
28	I have a lot of self-confidence when it comes to working with blogs.				

Outlier Impact and Accommodation Methods: Multiple Comparisons of Type I Error Rates

Hongjing Liao

Beijing Foreign Studies University
Beijing, China

Yanju Li

Western Carolina University
Cullowhee, NC

Gordon Brooks

Ohio University
Athens, OH

A Monte Carlo simulation study was conducted to examine outliers' influence on Type I error rates in ANOVA and Welch tests, and the effectiveness of two outlier accommodation methods: nonparametric rank based method and Winsorizing. Recommendations are given regarding outlier handling with different sample sizes and number of outliers.

Keywords: outliers, type I error, Monte Carlo simulation, outlier accommodation, nonparametric, Winsorizing

---

Introduction

Extreme data points, or outliers, requires attention and investigation (Barnett & Lewis, 1994). Outliers are often inevitably seen in data sets of educational research projects, even when data come from reputable sources and the data collection is carefully executed. The existence of outliers has been recognized and noted for centuries, and the outlier problem is generally seen as “reducing and distorting the information about the data source or generating mechanism” (Barnett & Lewis, 1994, p. 4). To put it in a statistical context, there are concerns about the disproportionate influence of outliers on statistical analyses, based on sample means and variance. Studies have provided evidence that shows the effect of outliers resulted in inflation of Type I error rates and reduced power in parametric t and F tests (Barnett & Lewis, 1994; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Wilcox, 1998; Zimmerman, 1994b).

Because distortions of statistical significance tests could lead to faulty conclusions if indications of outliers are not carefully examined, it is natural to seek a means of identifying and explaining outliers. A number of studies are devoted to

---

Hongjing Liao is a lecturer. Email her at: hongjing.liao@139.com. Yanju Li is a Research Development Specialist. Email her at: yanjuli@wcu.edu. Dr. Brooks is an Associate Professor of Educational Research and Evaluation.

investigating the sources of outliers and detecting their presence in various data sets and distributions (Beckman & Cook, 1983). However, very few give emphasis on how to handle outliers, and there are even fewer studies that compare different outlier accommodation techniques. It is useful to investigate the circumstances under which outliers can be treated, as well as the effectiveness of outlier treatment methods. Hence, the purpose of this study is to focus on the impact of outliers on significance tests, and presents simulation results for comparisons of outlier accommodation methods in order to provide recommendations for practice.

Outliers: Definition, Detection, and Accommodation

An outlier refers to an observation that “appears to be inconsistent with the remainder of that set of data” (Barnett & Lewis, 1994, p. 7). Although problems in statistical analyses caused by outliers are a concern in the development of statistical methods (Barnett & Lewis, 1994), perceptions about outliers evolve with the development of educational research methodologies. The restrictive view of outliers being erroneous and contaminating has changed. In a present perspective, outliers are an “empirical reality” (Rousseeuw & Van Zomeren, 1990, p. 650), and instead of being misleading and wrong, they could provide useful information about the sample and, in some situations, indicate that a different model or distribution may fit the data better (Barnett & Lewis, 1994).

In parametric analyses, outliers are often identified according to how particular data points deviate from the center (the mean) of the distribution of the data set. Thus, for a normally distributed data set, the common rule is that an outlier is any value that is beyond ± 3 standard deviations from the mean. In addition, for different research designs and methods of analysis, there are different approaches developed to detect outliers (Barnett & Lewis, 1994; Berkane & Bentler, 1988; Cook, 1986; Gnanadesikan, 1997; Jarrell, 1991). Some approaches are adapted from univariate methods, such as frequency tables, histograms, and box plots (Allison, Gorman, & Primavera, 1993; Jarrell, 1991); some use residuals of various kinds (Cook, 1986; David, 1978); others suggest bivariate and multivariate techniques such as Cook’s distance (Allison et al., 1993), principal components (Hawkins, 1974), hat matrix (Hoaglin & Welsch, 1978), and Mahalanobis distance (Stevens, 1984). However, with such a variety of approaches available, it is still the researcher’s decision to define outliers depending on research contexts, and researchers should always seek meaning and interpretation of outliers before rejecting or choosing any techniques to deal with the deviant observations. The

OUTLIER IMPACT AND ACCOMMODATION METHODS

reason for doing so is because, as the nature and origins of outliers differ, the approaches to handle outliers vary accordingly.

Outliers may arise for deterministic reasons or for less tangible reasons. Deterministic reasons refer to apparent errors in execution of data that are controllable and correctable. Examples of deterministic outliers include recording and calculating errors, erroneous data entries, and failure to specify missing values (Barnett & Lewis, 1994; Tabachnick & Fidell, 2001; Warner, 2008). For outliers that arise as a result of deterministic reasons, the remedy is simple and straightforward: to replace outliers with correct values. However, more often than not, the reasons for the existence of outliers are less clear-cut. Scholars suggested three major sources of outliers: inherent variability, measurement error, and execution error (Anscombe, 1960; Barnett, 1978; Grubbs, 1969; Hampel et al., 1986; Tabachnick & Fidell, 2001). First, inherent variability refers to the variations demonstrated by outliers as a natural feature of the population under study. In other words, the outlying observations are representative of the target population, because the population has more extreme scores than a normal distribution. Of course, outliers are also possible as part of a normal distribution. Second, the occurrence of outliers could also be due to measurement error, such as rounding errors, recording errors, or variability imposed due to an inadequate measuring instrument. Finally, an execution error could be another source of outlying observations, such as a biased sample that includes individuals who are not truly representative of the population. Although theoretically, measurement and execution errors could be examined and corrected, in many circumstances it is very difficult, or even impossible, in practical research projects to distinguish from which sources outliers truly rise.

For less tangible outliers, the reasons for their occurrences are often not clear; there are two basic approaches to handle such outliers: to reject the outliers or to retain and accommodate the outliers to reduce their effect (Jarrell, 1991; Warner, 2008). Rejection of outliers includes simple removal of outliers after taking into account the appropriateness of all data (Field, 2011). Or, as Allison et al. (1993) suggested, rejecting outliers can also include running the analyses with and without outliers, comparing the results and reporting an assessment of the influence of outliers through deletion.

However, it is often not encouraged to reject outliers, especially when there is no tangible explanation about the occurrence of outliers. Outliers can be legitimate data points and removal may cause loss of useful information (Orr, Sackett, & DuBois, 1991). Sometimes outliers may reflect unusual but substantively meaningful aspects of the intended study (Chow, Hamaker, & Allaire,

2009; Hampel, 2001). An alternative approach is to use accommodation methods to reduce the impact of the outlying observations, including utilizing robust tests and outlier treatment methods. However, even with outlier accommodation approaches effectively applied, it is uncertain that the influence of outliers can be removed completely, but the aim is to minimize such influence.

There are several approaches that can be used to diminish or lower the impact of outliers, such as log transformation (Warner, 2008), nonparametric statistical ranking (Zimmerman & Zumbo, 1990), and Winsorizing (Dixon & Tukey, 1969; Dixon & Yuen, 1974). Among statistical tests, nonparametric methods based on ranks are argued to effectively control Type I error rates in the presence of outliers (Zimmerman & Zumbo, 1990). For example, Zimmerman (1994b; 1995) reported that compared to parametric methods, the Mann-Whitney-Wilcoxon test can effectively control Type I error, and nonparametric methods based on ranks exhibited slightly better Type I error rates than ANOVA methods for several outlier-prone and non-normal distributions. Zimmerman and Zumbo (1990) showed that the Type I error rates of Mann-Whitney-Wilcoxon and pooled-variance Student t test were relatively equivalent under simple bounded transformations used to handle outlier-prone distributions.

Furthermore, Winsorizing is another popular method to reduce the weights of outliers by replacing them with a specific percentile of data-dependent values (Dixon & Yuen, 1974; Orr et al., 1991). In practice, the location of Winsorization often depends on prior knowledge, and is suggested to be adjusted according to the shapes of the distribution (Dixon & Yuen, 1974; Tukey, 1962). This study, therefore, explores the effectiveness of the Winsorizing approach with different percentiles.

Purpose of the Study

This study is primarily motivated by two very practical questions: what is the impact of outliers on Type I error rates with different sample sizes and number of outliers, and which outlier accommodation methods can effectively control Type I error rates under varying sample size and outlier number conditions? Therefore, the purpose and contributions of this study are three-fold:

First, this study started by examining outliers' influence on Type I error rates in ANOVA and Welch tests with different sample sizes and number of outliers, and further explored distinct features of such influence in various combinations of conditions. Outlier impact in previous studies is often treated as a type of violation of normality, and the number of outliers in data sets was not studied separately

(Zimmerman, 1994a; 1994b; 1995). This study highlighted the association of outlier number and its influence on Type I error rates. Moreover, given outliers' influence on sample variance, both ANOVA and Welch tests are included and compared to take into consideration the problem of unequal variance in mean comparison analyses.

Secondly, although outlier accommodation methods are available and effective for different conditions and distributions (Dixon & Tukey, 1969; Dixon & Yuen, 1974; Orr et al., 1991; Zimmerman, 1995), no comparison between the methods are provided in terms of the effectiveness of Type I error control under the same conditions. This study investigated two basic approaches in handling outliers and how effective they were in controlling Type I error rates. Specifically, the study compared the Type I error rates when outliers are removed and retained using nonparametric methods and Winsorizing. Comparing the sensitivity of nonparametric and Winsorizing methods on outlier impact not only fills the current gap about the two methods, but can also provide basic information for guidelines of the use of outlier treatment methods. Finally, our study was also conducted to explore the Winsorizing methods with different Winsorization percentiles because no consensus has been reached regarding Winsorization locations, and little information was provided on how to decide the locations in existing literature.

In short, this study ventured to explore some new areas on outlier impact and outlier treatment based on existing studies. From the research design perspective, when the occurrence of outliers cannot be traced, which frequently happens in statistical analyses of educational research, it is reasonable to retain the outliers but give less weight to their influence. Therefore, understanding the impact brought by the presence of outliers and choosing an appropriate method for outlier accommodation are critical for credible analysis and conclusion.

Methods

A Monte Carlo program was developed in the R language for data simulation and computation of statistical results for different outlier and accommodation conditions. As a useful approach to evaluate the quality of statistical procedures (in this case the Type I error rate), a Monte Carlo program allows sample data to be drawn with many iterations in simulation. Rejection rates of significant tests could be counted with many iterations, through which Type I error rate under the true null hypotheses would be obtained (Mooney, 1997). In addition, R as an open-source computer statistical package and programming language has built-in functions to perform the ANOVA F test and the nonparametric Kruskal-Wallis rank sum test.

The general procedures for this simulation study are as follows: first, as this study focuses on multiple comparisons of Type I errors, samples of varied sample sizes and varied number of outliers were drawn from the same univariate normally distributed data simulated using the built-in R function `rnorm`. For each condition, equal sample sizes were manipulated for three groups and a varied number of outliers are included in only one group (group three). Second, ANOVA and Welch tests were performed using the same group of simulated data for analysis with outliers excluded, outliers included but with no treatment, and outliers accommodated by the nonparametric test and the Winsorizing method. For each condition, 10,000 replications were conducted and Type I error rates for different conditions were computed. Finally, simulation and statistical results were analyzed to examine outlier impact on Type I error rates, as well as advantages and disadvantages of the outlier treatment techniques under different conditions. Details about data generation, outlier injection, replication, and analysis procedures are provided in the following sections.

Data Generation

The sample sizes ($n = 20, 40, 60, 80$, and 100) were manipulated in the way that the three groups for ANOVA test always had equal sample sizes with the outlier(s) being inserted into only one group. 200,000 normally distributed $N(0, 1)$ cases were generated using the function `rnorm` (sample size, mean, standard deviation). The generated population data were split into two data sets: data without outliers ($u - 3\sigma \leq x \leq u + 3\sigma$) and data with outliers ($x < u - 3\sigma$ and $x > u + 3\sigma$). Data for each group of a sample were randomly selected from these two data sets. The built-in R function `sample` was used to randomly sample the required number of observations from different data sets. The random selection procedures were performed in the following way: first, for the first two groups that contain no outliers, n points of data were randomly sampled from each data set ($u_1 - 3\sigma \leq x \leq u_1 + 3\sigma$) and ($u_2 - 3\sigma \leq x \leq u_2 + 3\sigma$), respectively. For the third group that has inserted outliers, n_{outliers} outliers were sampled from the data set ($x < u_3 - 3\sigma$ and $x > u_3 + 3\sigma$), and the absolute value of each was taken to ensure positive outliers. Then the rest of data for group three, $n - n_{\text{outliers}}$ ($n_{\text{outliers}} = 0, 1, 2, 3, 4$, and 5) number of data were sampled from the data set ($u_3 - 3\sigma \leq x \leq u_3 + 3\sigma$). To study Type I error rates, the null hypothesis is set to be true. Therefore, each group was randomly drawn from the same normal distribution $N(0, 1)$.

Outlier Injection

As indicated in the sampling procedures, outliers were sampled from data beyond 3 standard deviations on both directions of the generated data, and were injected into each sample. Generating data only between $u - 3\sigma$ and $u + 3\sigma$ results in a slightly decreased standard deviation than the population value but provides a certain gap between normal data and outliers. By “injecting” outliers into the normal data, we can ensure the required number of outliers for the research purpose. This differs from the approach adopted by many to use a “contaminated” standard normal distribution, where some data are generated $N(0, 1)$ while other data are generated at perhaps $N(0, 3)$ or $N(2, 1)$. The difference between the “injection” and the “contamination” methods lie in that the “injection method” guarantees that outliers are from the same normally distributed population and that they are included in every sample. The design is important to study Type I errors because the null hypothesis is held true when drawing the whole sample, including normal data and outliers from the same population.

Replication

10,000 replications were conducted for each condition to minimize the Monte Carlo sampling impact. Robey and Barcikowski (1992) tabulated the number of iterations required for examining departures from varied nominal Type I error rates. Mooney (1997) proposed that the more the better in choosing the number of iterations for Monte Carlo simulations. Thus, in order to sufficiently ensure the stability and generalizability of the results and, meanwhile, to avoid inefficiency in excessive iterations, 10,000 iterations were used for the current study.

Monte Carlo Analysis

For each sample from the simulated population (e.g., $u_1 = u_2 = u_3 = 0$, $n = 20$, $n_{\text{outliers}} = 1, 2, 3, 4, 5$, $sd = 1$), ANOVA and Welch tests were used to test the Null hypothesis. Statistical p values were documented for data with no outliers, data with outliers, data with outliers deleted, and data treated by two commonly used outlier accommodation methods: nonparametric and Winsorizing (Winsorized at 95th, 90th, 85th, 80th, and 75th percentile). In R codes, the function `anova` was used except for the nonparametric Kruskal-Wallis test, which used the built-in R function `kruskal.test`.

Type I error rates were calculated at the nominal significance level $\alpha = 0.05$. The calculated p values were compared to the liberal criterion $\alpha \pm 1/2\alpha$ with an interval of $[0.025, 0.075]$, the stringent criterion $\alpha \pm 1/10\alpha$ with an interval of $[0.045, 0.055]$ (Bradley, 1978), and the intermediate criterion $\alpha \pm 1/4\alpha$ with an interval of $[0.0375, 0.0625]$ (Robey & Barcikowski, 1992).

Table 1. Type I error rates of parametric significance tests and outlier removed under varied sample sizes and outlier conditions

Sample Size	Outlier Number	Parametric		Outlier Removed	
		Anova	Welch	Anova	Welch
N = 20	0 outliers	0.0492	0.0467	0.0492	0.0467
	1 outlier	0.0455	0.0459	0.0503	0.0479
	2 outliers	0.0846	0.0702	0.0504	0.0489
	3 outliers	0.1599	0.1182	0.0486	0.0475
	4 outliers	0.3002	0.1940	0.0486	0.0475
	5 outliers	0.4881	0.3168	0.0486	0.0471
N = 40	0 outliers	0.0528	0.0528	0.0528	0.0528
	1 outlier	0.0533	0.0513	0.0513	0.0514
	2 outliers	0.0767	0.0707	0.0530	0.0525
	3 outliers	0.1233	0.1038	0.0517	0.0520
	4 outliers	0.1920	0.1536	0.0523	0.0525
	5 outliers	0.2918	0.2222	0.0516	0.0512
N = 60	0 outliers	0.0497	0.0522	0.0497	0.0522
	1 outlier	0.0509	0.0518	0.0511	0.0512
	2 outliers	0.0659	0.0638	0.0516	0.0516
	3 outliers	0.0981	0.0890	0.0516	0.0518
	4 outliers	0.1480	0.1288	0.0512	0.0514
	5 outliers	0.2156	0.1803	0.0507	0.0509
N = 80	0 outliers	0.0546	0.0535	0.0546	0.0535
	1 outlier	0.0520	0.0514	0.0538	0.0530
	2 outliers	0.0647	0.0625	0.0545	0.0531
	3 outliers	0.0925	0.0844	0.0546	0.0532
	4 outliers	0.1319	0.1161	0.0538	0.0528
	5 outliers	0.1820	0.1571	0.0531	0.0535
N = 100	0 outliers	0.0489	0.0483	0.0489	0.0483
	1 outlier	0.0507	0.0489	0.0499	0.0492
	2 outliers	0.0590	0.0568	0.0494	0.0486
	3 outliers	0.0809	0.0761	0.0505	0.0498
	4 outliers	0.1104	0.1018	0.0501	0.0489
	5 outliers	0.1523	0.1338	0.0503	0.0495

OUTLIER IMPACT AND ACCOMMODATION METHODS

Table 2. Type I error rates of different outlier accommodation techniques under varied sample sizes and outlier conditions

Sample size	Outlier number	Nonparametric	Winsorizing									
			95th percentile		90th percentile		85th percentile		80th percentile		75th percentile	
			Anova	Welch	Anova	Welch	Anova	Welch	Anova	Welch	Anova	Welch
N = 20	0 outliers	0.0480	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467	0.0492	0.0467
	1 outlier	0.0459	0.0514	0.0504	0.0539	0.0523	0.0536	0.0533	0.0540	0.0542	0.0551	0.0534
	2 outliers	0.0583	0.0832	0.0700	0.0711	0.0663	0.0675	0.0645	0.0654	0.0635	0.0654	0.0642
	3 outliers	0.0873	0.1601	0.1182	0.1535	0.1167	0.1090	0.0974	0.0928	0.0870	0.0850	0.0801
	4 outliers	0.1348	0.3081	0.1973	0.2912	0.1917	0.2804	0.1885	0.1752	0.1470	0.1288	0.1197
	5 outliers	0.2098	0.5090	0.3259	0.4839	0.3174	0.4670	0.3115	0.4525	0.3052	0.2766	0.2245
N = 40	0 outliers	0.0507	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528	0.0528
	1 outlier	0.0508	0.0555	0.0533	0.0562	0.0544	0.0565	0.0545	0.0561	0.0544	0.0552	0.0539
	2 outliers	0.0593	0.0674	0.0646	0.0621	0.0610	0.0615	0.0609	0.0602	0.0600	0.0605	0.0602
	3 outliers	0.0748	0.1176	0.1019	0.0818	0.0780	0.0742	0.0710	0.0708	0.0676	0.0674	0.0654
	4 outliers	0.0988	0.1847	0.1509	0.1297	0.1172	0.0990	0.0956	0.0864	0.0847	0.0794	0.0766
	5 outliers	0.1298	0.2872	0.2217	0.2664	0.2124	0.1429	0.1322	0.1118	0.1073	0.0964	0.0939
N = 60	0 outliers	0.0508	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522	0.0497	0.0522
	1 outlier	0.0511	0.0518	0.0527	0.0520	0.0531	0.0519	0.0528	0.0520	0.0528	0.0527	0.0531
	2 outliers	0.0559	0.0593	0.0596	0.0562	0.0572	0.0550	0.0558	0.0550	0.0558	0.0553	0.0551
	3 outliers	0.0644	0.0780	0.0745	0.0682	0.0679	0.0626	0.0624	0.0602	0.0590	0.0591	0.0586
	4 outliers	0.0805	0.1379	0.1215	0.0872	0.0839	0.0773	0.0744	0.0684	0.0689	0.0646	0.0650
	5 outliers	0.1004	0.2049	0.1735	0.1221	0.1146	0.0940	0.0920	0.0812	0.0806	0.0729	0.0732
N = 80	0 outliers	0.0514	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535	0.0546	0.0535
	1 outlier	0.0511	0.0554	0.0537	0.0549	0.0542	0.0550	0.0540	0.0550	0.0536	0.0551	0.0531
	2 outliers	0.0548	0.0591	0.0599	0.0586	0.0574	0.0576	0.0566	0.0566	0.0563	0.0566	0.0557
	3 outliers	0.0620	0.0730	0.0695	0.0654	0.0643	0.0627	0.0619	0.0599	0.0600	0.0594	0.0591
	4 outliers	0.0742	0.1002	0.0932	0.0765	0.0756	0.0721	0.0700	0.0677	0.0661	0.0631	0.0621
	5 outliers	0.0913	0.1674	0.1483	0.0992	0.0942	0.0820	0.0802	0.0758	0.0749	0.0718	0.0693
N = 100	0 outliers	0.0509	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483	0.0489	0.0483
	1 outlier	0.0513	0.0507	0.0501	0.0509	0.0504	0.0511	0.0501	0.0512	0.0504	0.0516	0.0499
	2 outliers	0.0531	0.0542	0.0538	0.0531	0.0534	0.0532	0.0525	0.0530	0.0517	0.0522	0.0511
	3 outliers	0.0606	0.0632	0.0619	0.0593	0.0591	0.0575	0.0569	0.0560	0.0551	0.0556	0.0540
	4 outliers	0.0697	0.0823	0.0799	0.0689	0.0679	0.0633	0.0635	0.0598	0.0606	0.0591	0.0580
	5 outliers	0.0795	0.1125	0.1052	0.0818	0.0800	0.0719	0.0715	0.0658	0.0658	0.0628	0.0622

Results

The Monte Carlo simulation results are summarized in [Table 1](#) and [Table 2](#). The results include Type I error rates of parametric significance tests and different outlier accommodation techniques under five sample sizes (20, 40, 60, 80, and 100) with six outlier conditions (outlier = 0, 1, 2, 3, 4, 5). Each entry in the tables is the probability of falsely rejecting the null hypothesis under the situation of a true Null (the Type I error rate). The rows represent sample sizes and the number of outliers, and the columns represent the significance tests with either untreated or treated outliers.

The Influence of Outliers on Statistical Results

The first two major columns in [Table 1](#), “parametric” and “outlier removed” columns, provide the Type I error rates in ANOVA and Welch tests for untreated outliers and after outliers being removed from the data set. The results show a clear influence of outliers on the statistical results of significance tests, which is illustrated by inflated Type I error rates. When outliers are removed from the data set, the Type I error rates of both ANOVA and Welch tests drop back to a significance level of around 0.05. Results in the parametric column show the general trend that with an increasing number of outliers being “injected” into the sample, the probability of Type I error increases from the significance level of 0.05 to a maximum of 0.4881 (ANOVA, $n = 20$, $n_{\text{outliers}} = 5$).

There are several notable features in the parametric test results regarding the influence of outliers on statistical results. First, the impact of outliers reflected in inflated Type I error rates varies significantly depending on the number of outliers present in the data set. [Table 1](#) shows the number of outlier conditions from 0 to 5. [Figure 1](#) shows the Type I error rates of ANOVA and Welch tests when no outlier and only one outlier is present. As it is shown in [Figure 1](#), when there is only one outlier, the Type I error rates maintain around the significance level 0.05 regardless of the sample size and significance tests. For both ANOVA and Welch tests, a single outlier exerts little modification on the false rejection rates. Compared with a single outlier in the data set, there is an apparent inflation of Type I error rates when there are two outliers.

For example, when there are two outliers, the Type I error rates of ANOVA tests for sample size 20, 40, 60 and 80 are 0.0846, 0.0767, 0.0659 and 0.0647, respectively. All of them exceed the upper bound of Robey and Barcikowski’s

OUTLIER IMPACT AND ACCOMMODATION METHODS

(1992) intermediate criterion 0.0625. When there are three outliers or more, there is an even more dramatic increase in the Type I error rate across all sample sizes, all of which become greater than the upper bound of the liberal criterion 0.075 (Bradley, 1978), with the lowest Type I error rates as 0.0761 (Welch, $n = 100$) and 0.0809 (ANOVA, $n = 100$). This tendency of inflated Type I errors with an increased number of outliers can be clearly shown in the graphical representations of Figure 2, the Type I error rates of ANOVA and Welch tests with 0 to 5 number of outliers across five different sample sizes.

In addition to the number of outliers casting an impact on the Type I error rate, the second feature involves sample sizes. As it can also be shown in Figure 2, the magnitude of Type I error rate inflation decreases with the growth of sample size. In other words, the impact of outliers on the false rejection rates is substantially greater with smaller sample sizes, and as sample size increases, the impact of outliers decreases although it is still inflated.

When other conditions hold the same, the Welch test showed a better control of Type I error rates when compared with the ANOVA test in presence of outliers. Although the Type I error rates are inflated beyond Bradley's (1978) criteria for both ANOVA and Welch tests when there are more than three outliers, at each sample size level with the same number of outliers, the Welch test has a less inflated Type I error rates than the ANOVA test.

Figure 1. Type I error rates for ANOVA and Welch when zero and one outlier exist

Figure 2. Type I error rates for ANOVA and Welch with varied sample sizes and number of outliers.

Outlier Treatments Methods can Reduce Outlier Influence on Statistical Results

Table 2 provides Type I error rates of significance tests with outlier accommodation methods being applied: the robust approach of using nonparametric Kruskal-Wallis test and the outlier treatment method of Winsorizing at the 95th, 90th, 85th, 80th, and 75th percentiles.

Overall, both the nonparametric Kruskal-Wallis test and the Winsorizing method are effective in reducing outlier influence on statistical results. Figure 3 shows graphical comparisons of Type I error rates of ANOVA and Welch with untreated and treated outliers under various conditions. For sample size equal to 20, 40, 60, 80 and 100, the outlier treatment method of Winsorizing is illustrated at 75th, 80th, 85th, 85th and 90th percentiles as examples, at which the Type I error rates are acceptable. It can be seen from Figure 3, with untreated outliers, the Type I error rates inflate rapidly as the number of outliers increase. Comparatively, the Type I error rate inflation is reduced to the acceptable intervals of criteria when outlier accommodation methods are used. This tendency in the results is not only accurate for the examples in Figure 3; it is also consistent across all sample size and methods conditions.

OUTLIER IMPACT AND ACCOMMODATION METHODS

Figure 3. Type I error rates for ANOVA, Welch, nonparametric, and (a) 75 percentile of Winsorizing with sample size of 20; (b) 80 percentile of Winsorizing with sample size of 40; (c) 85 percentile of Winsorizing with sample size of 60; (d) 85 percentile of Winsorizing with sample size of 80; (e) 90 percentile of Winsorizing with sample size of 100.

Outlier Accommodation Methods: Sensitivity

To a certain extent, the outlier treatment methods “corrected” the influence of outliers on the statistical results, although the degree of correction varies for different methods. The two outlier accommodation techniques perform differently in minimizing the impact of outliers under different conditions.

Figure 4. Type I error rates for nonparametric with varied sample size and number of outliers.

The effectiveness of the Kruskal-Wallis test in controlling Type I error rates depends jointly on the number of outliers and sample size. The intertwining effect of sample size and outlier numbers can be observed in Figure 4, which shows the Type I error rates across five different sample sizes of the nonparametric Kruskal-Wallis test in the presence of a different number of outliers.

First, with respect to the number of outliers, two or fewer outliers show little modification in the probability of Type I error rates for the Kruskal-Wallis test, and this result is in accord with conclusions of previous studies about the robustness of nonparametric tests under violations of normality (e.g., Zimmerman, 1994b; 1995). However, when there are three or more outliers present, there is still discernable inflation of Type I errors, and the Kruskal-Wallis test is not able to effectively control Type I error rates to be within the interval of Bradley's (1978) standards.

Second, similar to the impact of untreated outliers on the probability of Type I errors, sample size plays a role regarding the magnitude of change: the larger the sample size, the relatively less inflation in Type I error rates.

Table 2 shows the Type I error rates of ANOVA and Welch tests after outliers being Winsorized at five different percentiles under varying sample size and outlier number conditions. In other words, the injected outliers in each data set are replaced by the scores at the assigned percentile (95th, 90th, 85th, 80th, and 75th). Similarly to

the nonparametric test, Winsorizing also shows an effective control of Type I error rates, and the effectiveness of Winsorizing varies at different Winsorization locations. How much to Winsorize in order for a reasonable control of Type I error is jointly affected by the sample size and the number of outliers.

A smaller Winsorization percentile, such as the 75th or 80th percentile, is necessary to control Type I error rates when sample sizes are small. As the sample size increases, the impact of outliers on probability of Type I errors decreases, and a relatively larger percentile (90th or 95th) of Winsorization is sufficient to accommodate the effects of outliers to achieve an acceptable Type I error rate. Regarding the number of outliers, with two or fewer outliers, Winsorizing at the 95th percentile shows a good control of Type I error rates across all sample sizes except when $n = 20$. At each sample size, with growing number of outliers in the data set, a smaller percentile is necessary for a good control of Type I error rates. However, it is important to note that when sample sizes are small, such as $n = 20$ and 40, even Winsorization at the 75th percentile does not show very effective control of Type I error rates when there are four or more outliers. As sample sizes grow bigger, a 75th percentile Winsorizing can reduce the inflated Type I error to meet the intermediate or liberal standards (Bradley, 1978; Robey & Barcikowski, 1992).

Conclusion

Based on the results and figures presented in the result section, certain statements of existing studies regarding the impact of outliers were replicated. In addition, it was confirmed that outliers can change the probability of Type I errors by exerting a disproportionate influence on means and variances in parametric F tests such as ANOVA and Welch tests. The current study provides new evidence in two ways: first, positive outliers inserted into one of the three groups can inflate the Type I error rates of F tests when the null hypothesis is true. Secondly, for previous studies, the impact of outliers was investigated using the contamination method in an outlier-prone data set (Zimmerman, 1994a; 1994b; 1995), in which the precise number of outliers or the extremity of outliers are not specified at each condition. The current study, by adopting the injection method, specified the number and relative extremity of the outliers, and made sure that the inserted outliers did belong to the population. The current study comes to similar conclusions with studies using contamination methods, and further confirms the impact of outliers under a different circumstance.

Furthermore, regarding outlier impact on a nonparametric test, different from conclusions drawn from mixed-normal distributions where outliers were studied in Zimmerman's (1994b; 1995) studies, the current study investigated different outlier number and sample size conditions, and presented similar results under certain conditions and different conclusions in other conditions. The impact of outliers on nonparametric tests in terms of Type I error rates depends on sample size and the number of outliers. When sample size is relatively large ($n = 80$ and 100), a nonparametric test has a good control of Type I error even when there are five outliers, which is in accord with the results of previous studies (Zimmerman, 1994b; 1995). However, when the sample size is small, there is non-ignorable inflation of Type I error caused by outlier influence, especially with two and more outliers present. Therefore, the nonparametric test is robust against outlier influence, but more attention should be paid when the sample size is small.

It is the number of outliers that seems to matter on the issue of outlier impact on the statistical results, regardless of the sample size. In other words, no matter how large the sample size is, the false rejection rates almost adhere to the nominal significance level (0.05) when the number of outliers is less than two, indicating that no accommodation techniques are necessary. As the number of outliers increases, the inflation of Type I errors begins to appear. Different outlier accommodation techniques have similar effect when the number of outliers was less than two, but the effect began to differ greatly as the number of outliers increased.

As for the comparison of sensitivity to outlier influence between nonparametric test and Winsorizing, it largely depends on the number of outliers in the data set and the location of Winsorization. When there are only two outliers, both the nonparametric test and Winsorizing methods show an effective control of Type I error rates. Yet, when sample sizes are small, the nonparametric test shows a better control of Type I errors than Winsorizing at the 95th and 90th percentile, but the two accommodations methods yield similar results at the 85th and 80th percentile of Winsorization. Therefore, with relatively small sample sizes, nonparametric could have an advantage over Winsorizing in controlling Type I error rates, especially when a large Winsorization percentile is preferred. When there are more than three outliers, a nonparametric test is still relatively robust with large sample sizes, but it does not show a good control of Type I errors when sample sizes are as small as 20 and 40. Comparatively, the Winsorizing with different percentiles can still maintain a good control of the probability of Type I error except that, as the number of outliers increases, the Winsorization location requires a smaller percentile. Thus, when encountering a relatively small sample size with three or

OUTLIER IMPACT AND ACCOMMODATION METHODS

more outliers present, the Winsorizing method offers a more accurate control of Type I error rates than the nonparametric test.

Recommendations can be made regarding the choices of parametric tests when outliers exist in the data and outlier accommodation methods. First, since a single outlier can have little impact on the Type I error rates under a true Null condition, researchers can keep the outlier in the data regardless of the sample size and the Type of F tests applied. However, if there is more than one outlier, the Welch test shows a better performance in controlling Type I error rates and is therefore recommended over ANOVA. Second, with regard to the choice of outlier accommodation methods, the nonparametric test is recommended for small sample sizes when two or less outliers are identified, or for large sample sizes when the number of outliers exceeds three. In addition, the method of Winsorizing is able to accommodate different sample size and outlier number conditions with different Winsorization percentiles. The smaller the sample size and the more outliers, the smaller percentile of Winsorization is required to have a better control of Type I error rates.

Many factors contribute to what approaches or methods should be taken in actual research, and the recommendations made in this study are solely based on the factor of controlling Type I error inflation and on the premise that both parametric and nonparametric approaches are available for use. It is recommended that the outliers be investigated as part of the research design before applying any accommodation techniques, and decisions on the choice of methods should consider the research design and methodology. Apart from Type I errors, other statistical factors such as power will also contribute to the effectiveness of outlier accommodation methods, which should be investigated in future studies of this topic.

References

- Allison, D. B., Gorman, B. S., Primavera, L. H. (1993). Some of the most common questions asked of statistical consultants: Our favorite responses and recommended readings. *Genetic, Social, and General Psychology Monographs*, 119(2), 155-184.
- Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, 2(2), 123-147. doi: [10.1080/00401706.1960.10489888](https://doi.org/10.1080/00401706.1960.10489888)
- Barnett, V. (1978). Convenient probability plotting positions for the normal distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 25(1), 47-50. doi: [10.2307/2346518](https://doi.org/10.2307/2346518)
- Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). Chichester, UK: Wiley.
- Beckman, R., & Cook, R. D. (1983). Outlier.....s. *Technometrics*, 25(2), 119-149. doi: [10.2307/1268541](https://doi.org/10.2307/1268541)
- Berkane, M., & Bentler, P. M. (1988). Estimation of contamination parameters and identification of outliers in multivariate data. *Sociological Methods & Research*, 17(1), 55-64. doi: [10.1177/0049124188017001003](https://doi.org/10.1177/0049124188017001003)
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152. doi: [10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)
- Chow, S., Hamaker, E. L., & Allaire, J. C. (2009). Using innovative outliers to detect discrete shifts in dynamics in group-based state-space models. *Multivariate Behavioral Research*, 44(4), 465-496. doi: [10.1080/00273170903103324](https://doi.org/10.1080/00273170903103324)
- Cook, R. D. (1986). Assessment of local influence. *Journal of the Royal Statistical Society. Series B (Methodological)*, 48(2), 133-169. Retrieved from <http://www.jstor.org/stable/2345711>
- David, H. A. (Ed.). (1978). *Contributions to survey sampling and applied statistics: Papers in honor of H. O. Hartley*. New York, NY: Academic Press.
- Dixon, W. J., & Tukey, J. W. (1969). Approximate behavior of the distribution of Winsorized t (trimming/Winsorization 2). *Technometrics*, 10(1), 83-98. doi: [10.1080/00401706.1968.10490537](https://doi.org/10.1080/00401706.1968.10490537)
- Dixon, W. J., & Yuen, K. K. (1974). Trimming and Winsorization: A review. *Statistische Hefte*, 15(2-3), 150-170. doi: [10.1007/BF02922904](https://doi.org/10.1007/BF02922904)

OUTLIER IMPACT AND ACCOMMODATION METHODS

Field, M. S. (2011). Application of robust statistical methods to background tracer data characterized by outliers and left-censored data. *Water Research*, 45(10), 3017-3118. doi: [10.1016/j.watres.2011.03.018](https://doi.org/10.1016/j.watres.2011.03.018)

Gnanadesikan, R. (1997). *Methods for statistical data analysis of multivariate observations*. (2nd ed.). New York, NY: Wiley.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1), 1-21. doi: [10.1080/00401706.1969.10490657](https://doi.org/10.1080/00401706.1969.10490657)

Hampel, F. R. (2001). Robust statistics: A brief introduction and overview. *Robust Statistics and Fuzzy Techniques in Geodesy and GIS Symposium*. Retrieved from <ftp://ess.r-project.org/Research-Reports/94.pdf>

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics: The approach based on influence functions*. New York, NY: Wiley.

Hawkins, D. M. (1974). The detection of errors in multivariate data using principal components. *Journal of the American Statistical Association*, 69(346), 340-344. doi: [10.1080/01621459.1974.10482950](https://doi.org/10.1080/01621459.1974.10482950)

Hoaglin, D. C., & Welsch, R. E. (1978). The hat matrix in regression and ANOVA. *The American Statistician*, 32(1), 17-22. doi: [10.1080/00031305.1978.10479237](https://doi.org/10.1080/00031305.1978.10479237)

Jarrell, M. G. (1991). Multivariate outliers: Review of the literature. *Annual Meeting of the Mid-South Educational Research Association*. Available from <http://eric.ed.gov/?id=ED339754>

Mooney, C. Z. (1997). *Monte Carlo Simulation*. Thousand Oaks, CA: Sage Publications.

Orr, J. M., Sackett, P. R., & DuBois, C. L. Z. (1991). Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology*, 44(3), 473- 486. doi: [10.1111/j.1744-6570.1991.tb02401.x](https://doi.org/10.1111/j.1744-6570.1991.tb02401.x)

Robey, R. R., & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology*, 45(2), 283-288. doi: [10.1111/j.2044-8317.1992.tb00993.x](https://doi.org/10.1111/j.2044-8317.1992.tb00993.x)

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633-639. doi: [10.2307/2289995](https://doi.org/10.2307/2289995)

- Stevens, J. P. (1984). Outliers and influential data points in regression analysis. *Psychological Bulletin*, 95(2), 334-344. doi: 10.1037/0033-2909.95.2.334
- Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). New York, NY: Allyn and Bacon.
- Tukey, J. W. (1962). The future of data analysis. *The Annals Mathematical Statistics*, 33(1), 1-67. Retrieved from <http://www.jstor.org/stable/2237638>
- Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniques*. Thousand Oaks, CA: SAGE Publication.
- Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, 53(3), 300-314. doi: 10.1037/0003-066X.53.3.300
- Zimmerman, D. W. (1994a). Increasing the power of the ANOVA F test for outlier-prone distributions by modified ranking methods. *Journal of General Psychology*, 122(1), 83-94. doi: 10.1080/00221309.1995.9921224
- Zimmerman, D. W. (1994b). A note on the influence of outliers on parametric and nonparametric tests. *The Journal of General Psychology*, 121(4), 391-401. doi: 10.1080/00221309.1994.9921213
- Zimmerman, D. W. (1995). Increasing the power of nonparametric tests by detecting and downweighting outliers. *Journal of Experimental Education*, 64(1), 71-85. doi: 10.1080/00220973.1995.9943796
- Zimmerman, D. W., & Zumbo, B. D. (1990). The relative power of the Wilcoxon-Mann-Whitney test and student *t* test under simple bounded transformations. *The Journal of General Psychology*, 117(4), 425-436. doi: 10.1080/00221309.1990.9921148

Z and t Distributions in Hypothesis Testing: Unequal Division of Type I Risk

Ceyhun O. Ozgur
Valparaiso University
Valparaiso, Indiana

Introductory statistics texts have given extensive coverage to two-sided inferences. All texts that were surveyed give significant coverage to one-sided hypothesis tests. Very few discussed the possibility of one-sided interval estimation at all. Even fewer mentioned so in any detail the possibility of dividing the risk of a type I error unequally between the tails for a two-sided confidence interval. None of the textbooks that were reviewed even considered the possibility of unequal tails for two-sided hypothesis tests. In this paper, we suggest that all statistics courses and texts should cover both one-sided tests and confidence intervals. Furthermore, coverage should also be given to unequal division of the nominal risk of a type I error for both hypothesis tests and confidence intervals. Examples are provided for both situations.

Keywords: Statistical Inference, Division of Type I error risk, Z distribution, Hypothesis testing, t distribution, Unequal Division of Type I Error

---

Introduction

One-sample Z and t tests are far less robust to departures from normality than are two sample tests. This has been known for over a century now. Hence, these techniques are usually taught only for pedagogical purposes. All introductory statistics textbooks cover two-sided hypothesis tests and confidence intervals with alpha split equally between the tails. All gave extensive coverage to one-sided hypothesis tests but only a few mention one-sided confidence intervals. Our research surveyed some of the most reputable textbooks used in Introductory Business Statistics courses. Only one textbook that was surveyed considered the possibility of splitting alpha unequally between the tails for a confidence interval and none did so for hypothesis tests. Since there is a one-to-one correspondence between hypothesis tests and confidence intervals, a strong case can be made that

---

Dr. Ozgur is a Professor of Information and Decision Science. Email him at: Ceyhun.Ozgur@valpo.edu.

all texts should give detailed coverage to one-sided confidence intervals. Practical examples are easy to develop and the intervals are simple to construct. Therefore, there is no justification for omitting their coverage.

In situations where both tails were actually of interest, all of the textbooks surveyed split alpha equally between the tails in hypothesis testing situations. Ramsey (1990) mentioned in his paper that instead of running a one-tailed test at the .05 level, the test could be ran at the .04 level. This would be more powerful than a two-tailed test ran at the .05 level because, for results in the predicted direction, such a test would be equivalent to a one-tailed test at the .025 level. However, results which come out in the opposite direction beyond a .01 probability could be rejected and taken as evidence against any previous knowledge about an outcome in the predicted direction. Of all the texts that were surveyed, only Harnett and Soni (1991) mentioned the possibility of an unequal split for confidence intervals and they only do so in one sentence. Students would develop a better understanding of the rationale underlying the choice of alpha if they were given a broad spectrum of possibilities for splitting it between the tails. This is particularly true for business students when examples can associate specific costs with type I errors. In practice, the costs associated with a type I error on one side may be different from the costs on the other side. Therefore, the risks should have been split proportionately to the costs. In the next section two examples will be provided to demonstrate the unequal split of α between the two tails. The example applications involve a service and a manufacturing scenario respectively. We surveyed many textbooks and the table of what we found regarding the uneven division of alpha can be found in the Table 1.

Service Application

Suppose that a fast food restaurant with significant sales from coffee customers at the drive-thru window is analyzing coffee temperature. The target temperature for a coffee cup is 175 degrees Fahrenheit with a temperature tolerance of ± 5 degrees. If a cup of coffee is too hot or in other words, if it is warmer than 180 degrees (above the upper tolerance) and someone gets burned, there is the potential for a very costly lawsuit, possibly a settlement for \$2.4 million. On the other hand, if the cup temperature is too cool, below the lower tolerance of 170 degrees, then there is likely to be lost sales and possible customer erosion in other product areas. The analyst estimates lost profits from declining sales to be \$800,000. In this scenario neither a one-tailed inference nor a two-sided inference with equal tails would be logically appropriate.

UNEQUAL DIVISION OF TYPE I RISK

Table 1. Statistical books with coverage of Z & t tests and unequal division of alpha

Authors	Publishers	Publication Date/Edition	Book Title	Z & t Test	Unequal Division of Alpha
A.D. Aczel	McGraw-Hill-Irwin	1999, 4 th edition	Complete Business Statistics	Yes	No
D. R. Anderson, D.J. Sweeney, & T. Williams	Southwestern	2009, 10 th edition	Statistics for Business and Economics	Yes	No
D. R. Anderson, D.J. Sweeney, & T. Williams	Southwestern	1998, 7 th edition	Quantitative Methods for Business	Yes	No
M. Berenson, & D.M. Levine	Prentice-Hall	1999, 7 th edition	Basic Business Statistics	Yes	No
M. Berenson, D. Levine, & T.C. Krehbiel	Prentice-Hall	2000, 2 nd edition	Business Statistics: A First Course	Yes	No
K. Black, & D. Eldredge	Southwestern	2002, 1 st edition	Business and Economic Statistics	Yes	No
B. Bowerman & R. O'Connell, & J.B. Orris	McGraw-Hill-Irwin	2004, 1 st edition	Essentials of Business Statistics	Yes	No
Y. Chou	Elsevier	1989, 1 st edition	Statistical Analysis for Business and Economics	Yes	No
W. Cochran, & G. Cox	John Wiley & Sons	1992, 2 nd edition	Experimental Designs	t-test: Yes Z-test: No	No
D.P. Doane, & L.E. Seward	McGraw-Hill	2016, 5 th edition	Applied Statistics in Business and Economics	Yes	No
D.L. Harnett & A.K. Soni	Addison-Wesley	1991, 1 st edition	Statistical Methods for Business and Economics	Yes	Yes
J. Hawkes, & W. Marsh	Hawkes Publishing	2005, 2 nd edition	Discovering Statistics	Yes	No
D.H. Hildebrand & R.L. Ott	Brooks/Cole	1998, 4 th edition	Statistical Thinking for Managers	Yes	No
G. Keller, B. Warrack, & H. Bartel	Wadsworth	1998, 1 st edition	Statistics for Management and Economics	Yes	No
R.I. Levin, & D.S. Rubin	Prentice-Hall	1994, 6 th edition	Statistics for Management	Yes	No
D.M. Levine, T. C. Krehbiel, M.L. Berenson	Prentice-Hall	2000, 2 nd edition	Business Statistics	Yes	No

Table 1 continued on next page

Table 1 cont. Statistical books with coverage of Z & t tests and unequal division of alpha

Authors	Publishers	Publication Date/Edition	Book Title	Z & t Test	Unequal Division of Alpha
R.D. Mason, & D.A. Lind	Irwin	1993, 8 th edition	Statistical Techniques in Business and Economics	Yes	No
J.T. McClave, P.G. Benson, & T. Sincich	Prentice-Hall	2011, 11 th edition	Statistics for Business and Economics	Yes	No
G. Meek, H. Taylor, K. Dunning, & K. Klafehn	Allyn & Bacon	1987, 1 st edition	Business Statistics	Yes	No
G. Meek, & S. J. Turner	Houghton & Mifflin	1983, 1 st edition	Statistical Analysis for Business Decisions	Yes	No
W. Mendenhall & R. Beaver	PWS-Kent	1992, 3 rd edition	A Course in Business Statistics	Yes	No
J. Neter, M. Kutner, C. Nachtsheim, & W. Wasserman	McGraw-Hill	1996, 4 th edition	Applied Linear Statistical Models	Yes	No
J. Neter, W Wasserman, G.A. Whitmore	Allyn & Bacon	1993, 4 th edition	Applied Statistics	Yes	No
P. Newbold	Prentice-Hall	1991, 3 rd edition	Statistics for Business and Economics	Yes	No
P. Newbold, W.L. Carlson, & B. Thorne	Prentice-Hall	2007, 6 th edition	Statistics for Business and Economics	Yes	No
M. Pelosi, & T. Sandifer	John Wiley & Sons	2002, 2 nd edition	Doing Statistics for Business with Excel	Yes	No
H. Scheffe	John Wiley & Sons	1999, 1 st edition reprinted	The Analysis of Variance	No	No
A.F. Siegel	Irwin	1990, 1 st edition	Practical Business Statistics	Yes	No
T. Sincich	Kraus	1996, 5 th edition	Business Statistics by Example	Yes	No
C. Watson, P. Billingsley, D.J. Croft, & D. Huntsberger	Allyn & Bacon	1990, 4 th edition	Statistics for Management and Economics	Yes	No
R. M. Weiers	Duxbury	2005, 5 th edition	Introduction to Business Statistics	Yes	No

Note: Most textbooks that were surveyed covered Z and t tests, as well as 1 and 2 sided hypothesis testing.

UNEQUAL DIVISION OF TYPE I RISK

Manufacturing Application

Consider a manufacturing process in which the first operation is to cut stainless steel rods into lengths of two feet prior to machining and threading operations. The nominal length for each piece is two feet. Customer specifications allow ± 0.050 inches. If a part is oversized, that is, greater than 24.050, it can be “reworked” at a cost of \$0.50 while undersized pieces, those less than 23.950, must be scrapped at a cost of \$2.00. In this example the cost of a steel rod being longer than 24.05 is lower than the cost of a steel rod below the lower specification limit. This would be an example where dividing the risk of a type I error equally between the tails is not a reasonable choice nor is a one-sided approach.

Literature Review

As mentioned earlier, few authors discuss one-sided confidence intervals. Only one to my knowledge even considers the possibility of an asymmetric two-sided confidence interval for means. None considered dividing alpha unequally between the tails in two-sided hypothesis tests. Hildebrand and Ott (1998) presented a brief discussion of one-sided confidence intervals for a mean, providing an equation and a very brief example. Neter, Wasserman and Whitmore (1993) provided an optional section on one-sided confidence intervals with some development and examples. Meek and Turner (1983) provided a detailed example of one-sided intervals with cost considerations and discussion. The best and most detailed coverage of one-sided confidence intervals is in Siegel (1997). Siegel (1997) also included a discussion and examples of one-sided prediction intervals.

Of the texts that this paper surveyed, only Harnett and Soni (1991) mentioned the possibility of splitting alpha unequally in a two-sided situation. They provide a brief discussion about asymmetric two-sided confidence intervals and the related cost considerations; however, they then dismiss the idea on the basis that the costs are not easily obtained (Harnett & Soni, 1991). Of the thirty-one texts surveyed this was the only one that considered the possibility of an unequal split in two-sided hypothesis testing situations.

Asymmetric Two-Sided Confidence Intervals for a Population Mean

One-sided confidence intervals are very easy to construct. All textbooks covered one-sided tests of hypotheses. Thus, due to the correspondence between confidence intervals and hypotheses tests, it is simply proposed that all

introductory statistics texts should cover one-sided confidence intervals as well as one-sided tests of hypotheses. Coverage of one-sided confidence intervals can be used as an important tool to reinforce the explanation of one-sided tests of hypotheses since both one-sided hypothesis tests and the one sided confidence intervals involve the concept of allocating the risk of a type one error to only one tail.

It is proposed to restrict the discussion to the situation in which it is desirable, based on cost considerations, to construct a two-sided confidence interval with unequal tails. The development of the equations with the following remarks was prefaced. If a desired value is not contained in the confidence interval it is assumed that corrective action is to be taken. If the desired value for μ is below the lower limit of the confidence interval the mean would be adjusted downward. If the true mean is actually desired this will result in an increase of "small units". Correspondingly, if the desired value for μ is greater than the upper limit of the confidence interval, an attempt would be made to increase the mean, resulting in an increase of "large units" if the true mean is desired.

The mathematics of constructing such intervals for population means is quite simple. Once the decision has been made regarding how alpha is to be divided, the appropriate percentage points from either the standard normal or the t -table is selected. If the population standard deviation is known, the standard normal (Z) distribution should be used. However, if the population standard deviation is estimated, t distribution should be used. Assume that it has been decided to place $p\alpha$, $0 \leq p \leq 1$, in the upper tail and $(1 - p)\alpha$, $0 \leq p \leq 1$, in the lower tail. Assuming that σ is known, the resulting confidence limits are given by Equation 1.

$$\begin{aligned} \text{UL} &= \bar{X} + Z_{(p)\alpha} \frac{\sigma}{\sqrt{n}} \\ \text{and} \\ \text{LL} &= \bar{X} - Z_{(1-p)\alpha} \frac{\sigma}{\sqrt{n}} \end{aligned} \tag{1}$$

If the population standard deviation is not known, Equation 1 can be modified by replacing the Z -statistic with an appropriate t -statistic with $n - 1$ degrees of freedom, where n represents the sample size. If the population standard deviation is not known the resulting confidence limits are given by Equation 2.

UNEQUAL DIVISION OF TYPE I RISK

$$\begin{aligned} \text{UL} &= \bar{X} + t_{(p)\alpha, n-1} \frac{s}{\sqrt{n}} \\ \text{and} \\ \text{LL} &= \bar{X} - t_{(1-p)\alpha, n-1} \frac{s}{\sqrt{n}} \end{aligned} \tag{2}$$

If c_1 represents the cost associated with an unnecessary decrease in μ and c_2 is the cost associated with an unnecessary increase in μ , and then setting

$$p = \frac{c_1}{(c_1 + c_2)} \tag{3}$$

It will provide a split that equalizes the expected costs between the two tails. For c_1 equal to zero, or alternatively $c_2 = 0$, it becomes a one-sided interval, while $c_1 = c_2$ gives the standard two-sided symmetric confidence interval for a mean.

Service Application Revisited

Let us refer to the service example described earlier involving the temperature of coffee. In that situation, it could be stated that a type I error has occurred if the actual average temperature of cups of coffee is on target at 175 degrees, but a confidence interval indicates that the target value is not contained within the limits. There are two possible ways in which the interval may not contain the target value:

1. If the entire confidence interval is above 175 degrees, then the process mean would be adjusted downward. This type of miss is defined as the “low side” miss.
2. If the entire confidence interval falls below the target value of 175 degrees, then the process mean would be adjusted upward. This is defined as the “high side” miss.

Of course, both types of adjustments given above are erroneous. The “low side” miss would result in unnecessarily lowering the temperature, resulting in temperatures that are too cool. The “high side” miss would result in temperatures that are too hot. As mentioned earlier, if the average coffee temperature is too cool, then there is likely to be lost sales estimated at \$800,000. If the average

coffee temperature is too hot and a customer gets burned, the estimated cost of a lawsuit is \$2,400,000. Since an unnecessary downward adjustment is three times less costly than an unnecessary upward adjustment, it would be preferred to miss on the “low side” as opposed to missing on the “high side”. Since $c_1 = \$800,000$ and $c_2 = \$2,400,000$,

$$p = \frac{c_1}{(c_1 + c_2)} = \frac{800,000}{(800,000 + 2,400,000)} = \frac{1}{4} = .25; \quad 1 - p = .75$$

For illustrative purposes assume that σ is known and the population standard deviation is 1.2 degrees, the confidence level, $1 - \alpha$, is chosen to be 0.90, and a sample of 4 cups of coffee resulted in a mean temperature of 174.1 degrees. The appropriate Z statistic values are then determined from the normal Z table. Since $p(\alpha) = 0.25 * 0.10 = .025$ and $(1 - p)\alpha = 0.75 * 0.10 = .075$, the corresponding Z -values are $Z_{0.025} = 1.96$ and $Z_{0.075} = 1.44$ for the upper limit and lower limit respectively. The resulting confidence limits are calculated as follows:

$$UL = 174.1 + 1.96 \left(\frac{1.2}{\sqrt{4}} \right) = 175.276 \text{ degrees}$$

$$LL = 174.1 - 1.44 \left(\frac{1.2}{\sqrt{4}} \right) = 173.236 \text{ degrees}$$

If the traditional two-sided confidence interval is employed with an equal split of α between the two tails, using $Z_{0.05} = 1.645$ on both sides of the confidence interval, resulting in an upper limit and a lower limit of 175.09 and 173.11 respectively. If the estimated costs of c_1 and c_2 are accurate, then compared to the asymmetric limits, symmetric limits will not provide enough coverage on the upper tail and will provide unnecessarily high coverage on the lower tail. Equation 3 provides the expected cost of a type I error where $EC = \text{Expected Cost of type I error}$.

$$EC = p(\alpha)c_2 + (1 - p)(\alpha)(c_1) \quad (4)$$

Based on the asymmetric limits,

UNEQUAL DIVISION OF TYPE I RISK

$$EC = (.25)(.10)(2,400,000) + (.75)(.10)(800,000) = 60,000 + 60,000 = \$120,000.$$

On the other hand, based on the symmetric limits,

$$EC = (.5)(.10)(2,400,000) + (.5)(.10)(800,000) = 120,000 + 40,000 = \$160,000.$$

Therefore, if the symmetric confidence limits are used in lieu of asymmetric limits, the decision maker has to incur an additional expected cost of \$40,000 (160,000 - 120,000).

In this situation it is also possible to construct a one-sided confidence interval by constructing only the upper control limit or by constructing only the lower control limit. Since the cost of being above the upper control limit is greater than the cost of being below the lower control limit, the one-sided interval is employed by determining only the upper limit. In this case, $Z_{.10}$ equals 1.28 and the upper limit is calculated as follows:

$$UL = 174.1 + 1.28 \frac{1.2}{\sqrt{4}} = 174.868.$$

However, if the one-sided confidence interval is employed in lieu of asymmetric two-sided limits,

$$EC = (.10)(2,400,000) = \$240,000.$$

If the one-sided confidence limit is used in lieu of asymmetric limits, the decision-maker has to incur an additional expected cost of \$120,000 (240,000 - 120,000).

Production Application Revisited

This situation was chosen to involve stainless steel rods that were mentioned earlier. In that situation a "type I error" would be considered to have occurred if the process is set up correctly and is yielding average lengths at or very close to 24 inches, but a confidence interval indicates otherwise. If the interval does not contain the value 24 on the "low side", i.e., the entire interval is above 24, the process mean would be adjusted downward, while if it misses on the "high side" it would be adjusted upward. Either adjustment would actually be a mistake and would result in production of some discrepant parts. Adjusting downward would

result in some undersized parts at a cost of \$2.00 each while adjusting upward would produce some oversized parts that will need to be reworked at a cost of \$0.50 each. Since an unnecessary downward adjustment is four times as costly as an unnecessary upward adjustment, that would be preferred if it was to occur less often. In fact, in order to balance expected costs, it is preferred to make unnecessary downward adjustments only one quarter as often as unnecessary upward ones. Therefore, letting $c_1 = 2.00$ and $c_2 = 0.50$ gives

$$p = \frac{2.00}{(2.00 + .50)} 0.8, \text{ and } 1 - p = 0.2.$$

The following assumptions were made in order to construct the confidence interval. For illustrative purposes assume that σ is known to equal 0.020 inches, the confidence level, $(1 - \alpha)$, is chosen to be 0.95, and a sample of 16 items gave a mean of 24.008 inches. The next step was to determine the appropriate values from the normal Z table to be used.

$p(\alpha) = 0.8 * 0.05 = .04$ and $(1 - p)\alpha = 0.2 * 0.05 = .01$. The corresponding Z -values for the upper and lower limits respectively are $Z_{0.04} = 1.75$ and $Z_{0.01} = 2.326$. The resulting confidence limits are:

$$UL = 24.008 + 1.75 (0.020/4) = 24.0168 \text{ in.}$$

$$LL = 24.008 - 2.326(0.020/4) = 23.9964 \text{ in.}$$

There was no evidence to suggest that the mean is other than 24 and no adjustment would be made at this time.

If the traditional two-sided confidence interval with an equal split of alpha between the two tails used, $Z_{0.025} = 1.96$ is used on both sides of the confidence interval, resulting in an upper limit and lower limit of 24.0178 and 23.9982 respectively.

If assumed that c_1 and c_2 are appropriate cost estimates, then compared to the asymmetric limits, the symmetric limits result in unnecessarily high coverage of the upper tail and insufficient coverage of the lower tail. Utilizing Equation 3, the per part expected cost of a Type I error can be calculated based on the asymmetric limits as follows:

$$EC = (.8)(.05)(.50) + (.2)(.05)(2.0) = .02 + .02 = \$0.04$$

UNEQUAL DIVISION OF TYPE I RISK

On the other hand, based on the symmetric limits the per part expected cost is:

$$EC = (.5)(.05)(.50) + (.5)(.05)(2.00) = .0125 + .05 = \$0.0625$$

Therefore, if the symmetric confidence limits are used in lieu of the asymmetric limits, the decision maker has to incur an additional expected cost per part of \$.025 (.0625 - .04).

If the one-sided confidence interval was constructed by using only the lower control limit, $Z_{.05} = 1.645$ would be used and the lower limit is calculated as follows: $LL = 24.008 - 1.645 \frac{.02}{\sqrt{16}} = 23.9976$. However, if the one-sided confidence interval is utilized in lieu of asymmetric two-sided limits,

$$EC = (.05)(2.00) = \$0.10$$

If the one-sided confidence limit is used in lieu of asymmetric limits, the decision-maker has to incur an additional expected cost per part of \$.06 (.10 - .04).

Unequal Tails for Two-Sided Hypothesis Tests about a Population Mean

For two-sided hypothesis tests the approach is similar and one can use either the p-value approach with a nominal α or the usual comparison of a sample statistic (Z or t) to the appropriate critical value based on an unequal split of the nominal α between the two tails. Note that the confidence interval approach reacted to the position of μ_o relative to the sample mean while the hypothesis test approach reacted to the position of the sample mean relative to μ_o . Therefore, the decision rule regarding the split of α between the tails is exactly opposite for the two approaches. That is, if H_o is rejected incorrectly on the high side, the resulting action is to adjust the process downward. On the other hand, if H_o is rejected incorrectly on the low side the process would be adjusted upward.

Let c_1 represent the cost of an unnecessary decrease in μ , occurring when H_o is rejected on the high side, and c_2 represented the similar cost of an unnecessary increase in μ , occurring with a rejection on the low side. Then, for a given significance level of α it is supposed to put $(1 - p)\alpha$ in the upper tail for the test procedure and $p\alpha$ in the lower tail in order to balance the expected costs of a type I error between them. Thus, a two-sided test with unequal costs in the rejection areas, for the hypotheses $H_o: \mu = \mu_o$ versus $H_a: \mu \neq \mu_o$, assuming σ is known, one

would reject H_o if the calculated Z -value is either greater than $Z_{(1-p)\alpha}$ or less than $-Z_{p\alpha}$. Alternatively, H_o is rejected. If either

$$\bar{X} < \mu - Z_{p\alpha} \frac{\sigma}{\sqrt{n}} \text{ or } \bar{X} > \mu + Z_{(1-p)\alpha} \frac{\sigma}{\sqrt{n}} \quad (5)$$

Production Application Utilizing the same example as the one that was presented earlier when confidence intervals are discussed, the hypotheses: $H_o: \mu = 24$ versus $H_a: \mu \neq 24$ is stated. As before, the information includes the following: $n = 16$, $\sigma = 0.020$, $\alpha = 0.05$, $c_1 = 2.00$, $c_2 = 0.50$ and $\bar{X} = 24.008$, $p = 0.8$. Concluding that μ is less than 24 will result in adjusting the process upward, while concluding μ is greater than 24 results in a downward adjustment. Therefore, the null hypothesis would be rejected if either

$$\bar{X} > 24 + 2.326(.020/4) = 24.0116 \text{ or if}$$

$$\bar{X} < 24 - 1.750(.020/4) = 23.9913$$

Since the sample mean value of 24.008 is between the two numbers, H_o could not be rejected alternatively,

$$Z_{calc} = \frac{\bar{X} - \mu}{\frac{\sigma_X}{\sqrt{n}}} \quad (6)$$

$$Z_{calc} = \frac{24.008 - 24}{\frac{.02}{\sqrt{16}}} = \frac{.008}{.005} = 1.6$$

Since $1.6 < 2.326$ and $1.6 > -1.75$, H_o cannot be rejected and is concluded that the mean length of steel rods does not appear to differ from 24 inches.

Service Application Based on the service example described earlier involving the temperature of coffee, recall that $c_1 = \$800,000$, $c_2 = \$2,400,000$, $n = 4$, $\sigma = 1.20$, $\alpha = 0.10$, $\bar{X} = 174.1$, and $p = .25$. Let's assume that the two-sided hypotheses are stated as follows:

UNEQUAL DIVISION OF TYPE I RISK

$H_o: \mu = 175$ versus $H_a: \mu \neq 175$. Concluding that μ is less than 175 will result in adjusting the process upward, while concluding that μ is greater than 175 results in a downward adjustment of the temperature. Therefore, the null hypothesis would be rejected if either

$$\bar{X} > 175 + 1.44(1.2/2) = 175.864 \text{ or if}$$

$$\bar{X} < 175 - 1.96(1.2/2) = 173.824$$

Since the sample mean value of 174.1 is between the two numbers, H_o could not be rejected. Alternatively,

$$Z_{calc} = \frac{\bar{X} - \mu}{\frac{\sigma_x}{\sqrt{n}}} = \frac{174.1 - 175}{\frac{1.2}{\sqrt{4}}} = \frac{-0.90}{.6} = -1.5$$

Since $-1.5 < 1.44$ and $-1.5 > -1.96$, H_o is not rejected, and conclude that the average temperature of a cup of coffee is not significantly different than 175 degrees.

Future Research Considerations, Limitations and Conclusions

Most introductory statistics texts fail to discuss one-sided confidence intervals and, other than a brief discussion without an example by Harnett and Soni (1991), none even consider the possibility of splitting α unequally between the two tails when the costs associated with the tails are different. One-sided confidence intervals should be standard coverage just as one-sided hypothesis tests are. Also, I believe that students should be made aware that the division of the type I risk between the tails should be decided by the costs rather than by convenience. In addition to the unequal division of type I error risk between the two tails for a single mean, that could also utilize this approach when testing a single proportion or a single variance, as well as testing the difference between two parameters. The latter case, though, is difficult to envision in a practical situation. Unfortunately, in many instances the consequences of committing a type I error cannot easily be expressed in dollar terms. In certain instances the difficulty of quantifying the cost of a type I error is the major drawback of this approach. There are many instances

in which cost estimates of committing a type I error are not available or cannot be estimated. In other situations it is simply cannot be deciphered the cost of making a type I error above the upper control limit vs. cost of making a type I error below the lower control limit. In these situations, even though it is not ideal, it may still be better to utilize the asymmetric limits in lieu of the symmetric limits based on subjective considerations. However, the asymmetric limits and the unequal split of alpha work best when cost estimates (c_1 and c_2) are available and are reasonably accurate. A topic for further consideration is the impact of unequal tails on the type II error and on balancing expected costs with respect to it as well.

Acknowledgements

Special Thanks to Yang Li, Zachariah Hughes, and Taylor Colliau for their help in this research project.

References

- Aczel, A. D., & Sounderpandian J. (2002). *Complete business statistics* (5th ed.). New York, NY: McGraw-Hill/Irwin.
- Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2003a). *Essentials of statistics for business and economics* (3rd ed.). Mason, OH: Southwestern.
- Anderson, D. R., Sweeney, D. J., & Williams, T. (2003). *Modern business statistics with Microsoft Excel*. Mason, OH: Southwestern.
- Berenson, M. L., Levine, D. M., & Krehbiel, T. C. (2003). *Basic business statistics: Concepts and applications* (9th ed.). Upper Saddle River, NJ: Prentice Hall
- Black, K. (2001). *Business statistics: Contemporary decision making* (3rd ed.). Mason, OH: Southwestern.
- Bowerman, B. L., & O'Connell, R. T. (2007). *Business statistics in practice* (4th ed.). Boston, MA: McGraw-Hill/Irwin.
- Carlson, W. L., Thorne, B. (1997). *Applied statistical methods: For business, economics, and the social sciences*. Upper Saddle River, NJ: Prentice Hall.
- Chou, Y.-L. (1989). *Statistical analysis for business and economics*. New York, NY: Elsevier.
- Cochran, W. G., & Cox, G. M. (1957). *Experimental designs* (2nd ed.). New York, NY: J. Wiley & Sons.

UNEQUAL DIVISION OF TYPE I RISK

- Doane, D. P., & Seward, L. W. (2016). *Applied statistics in business and economics* (5th ed.). New York, NY: McGraw-Hill.
- Gitlow, H. S., Oppenheim, A. V., & Oppenheim, R. (1995). *Quality management: Tools and methods for the improvement* (2nd ed.). Burr Ridge, IL: Irwin.
- Gitlow, H. S., & Levine, D. M. (2005). *Six Sigma for green belts and champions: Foundations, DMAIC, tools, cases, and certification*. Upper Saddle River, NJ: Pearson Education.
- Harnett, D. L., & Soni, A. K. (1991). *Statistical methods for business and economics* (4th ed.). New York, NY: Addison-Wesley.
- Hildebrand, D. H., & Ott, R. L. (1998). *Statistical thinking for managers* (4th ed.). Pacific Grove, CA: Duxbury Press.
- Keller, G., Warrack, B., & Bartel, H. (1994). *Statistics for management and economics: A systematic approach* (3rd ed.). Belmont, CA: Duxbury Press.
- Kohler, H. (2002). *Statistics for business and economics*. Mason, OH: Southwestern.
- Levin, R. I., & Rubin, D. S. (1994). *Statistics for management* (6th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Levine, D. M., Stephan, D., Krehbiel, T. C., & Berenson, M. L. (2002). *Statistics for managers using Microsoft Excel* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
- Mason, R. D., & Lind, D. A. (1993). *Statistical techniques in business and economics* (8th ed.). Homewood, IL: Irwin.
- McClave, J. T., Benson, P. G., & Sincich T. (1998). *Statistics for business and economics* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Meek, G., Taylor, H., Dunning, K., & Klafehn, K. (1987). *Business statistics*. Boston, MA: Allyn & Bacon.
- Meek G. E & Turner, S. J. (1983) *Statistical analysis for business decisions*. Boston, MA: Houghton Mifflin.
- Mendenhall, W. & Beaver, R. J. (1992). *A course in business statistics* (3rd ed.). Boston, MA: PWS-Kent.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (2005). *Applied linear statistical models* (5th ed.). Boston, MA: McGraw-Hill.
- Neter, J., Wasserman, W., & Whitmore, G. A. (1993). *Applied statistics* (4th ed.). Boston, MA: Allyn and Bacon.

- Newbold, P. (1991). *Statistics for business and economics* (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.
- Pelosi, M., & Sandifer, T. (2002). *Doing statistics for business with Excel: Data, inference, and decision making* (2nd ed.). New York, NY: John Wiley and Sons, Inc.
- Ramsey, P. H. (1990). 'One-and-a-half-tailed' tests of significance. *Psychological Reports*, 66(2), 653-654. doi:[10.2466/pr0.1990.66.2.653](https://doi.org/10.2466/pr0.1990.66.2.653)
- Scheffé, H. (1959). *The analysis of variance*. New York, NY: J. Wiley & Sons.
- Siegel, A. F. (1997). *Practical business statistics* (4th ed.). Chicago, IL: Irwin.
- Sincich, T. (1996). *Business statistics by example* (5th ed.). Upper Saddle River, NJ: Prentice Hall.
- Taguchi, G. (1986). *Introduction to quality engineering*. White Plains, NY: Kraus International Publications.
- Watson, C., Billingsley, P., Croft, D. J., & Huntsberger, D. (1993). *Statistics for management and economics* (5th ed.). Boston, MA: Allyn and Bacon.
- Weiers, R. M. (2002). *Introduction to business statistics* (4th ed.). New York, NY: Duxbury.

Bayesian Estimation of $P[Y < X]$ Based on Record Values from the Lomax Distribution and MCMC Technique

Mohamed A. W. Mahmoud

Al-Azhar University
Cairo, Egypt

Rashad M. El-Sagheer

Al-Azhar University
Cairo, Egypt

Ahmed A. Soliman

Islamic University of Madinah
Medina, Saudi Arabia

Ahmed H. Abd Ellah

Sohag University
Sohag, Egypt

Our interest is in estimating the stress-strength reliability $R = P[Y < X]$, where X and Y follow the Lomax distribution with common scale parameter. We discuss the problem in the situation where the stress measurements and the strength measurements are both in terms of records. Firstly, we obtain the MLE of R in general case (the common scale parameter is unknown). The MLE of the three unknown parameters can be obtained by solving one non-linear equation. We provide a simple fixed point type algorithm to find the MLE. We propose percentile bootstrap confidence intervals of R . A Bayes point estimator of R , and the corresponding credible interval using the MCMC sampling technique have been proposed. Secondly, assuming the common scale parameter is known, the MLE of R is obtained. Using exact distributions of the MLEs of the two unknown parameters, we construct the exact confidence interval of R . In this case, Bayes estimators have been obtained using Lindley's approximations. Analysis of a simulated data set has been presented for illustrative purposes. Finally, the different proposed methods have been compared via Monte Carlo simulation study.

Keywords: Stress-strength model, Lomax distribution, maximum likelihood estimation, bootstrap confidence intervals, credible intervals, Gibbs sampling, Markov chain Monte Carlo

---

Mohamad Mahmoud is Faculty of Science in the Mathematics Department. Dr. Rashad Mohamed El-Sagheer is a lecturer of Mathematical Statistics in the Mathematics Department, Faculty of Science. Email at rashadmath@azhar.edu.eg, rashadmath@yahoo.com. Ahmed Soliman and Ahmed Abd Ellah are Faculty of Science in their respective Mathematics Departments.

Introduction

The problem of estimating $R = P[Y < X]$ arises in the context of mechanical reliability of a system with strength (or supply) X and stress (or demand) Y , and R is chosen as a measure of system reliability. The system fails if and only if at any time the applied stress is greater than its strength. This type of reliability model is known as the stress-strength model. This problem also arises in situations where X and Y represent lifetimes of two devices and one wants to estimate the probability that one fails before the other. For example, in biometrical studies, the random variable X may represent the remaining lifetime of a patient treated with a certain drug, and Y represent the remaining lifetime when treated by another drug.

Review of literature

Parametric and nonparametric inferences on R for several specific distributions of X and Y under different conditions have been found in the literature. Nadarajah (2004a; 2004b) estimated $R = P[Y < X]$ from Logistic and Laplace distributions. Kundu and Gupta (2005) derived the maximum likelihood estimator of R and its asymptotic distribution when X and Y are independently distributed as generalized exponential distribution. Surles and Padgett (2001) considered the estimation of R where X and Y are Burr-X random variables. The theoretical and practical results on the theory and applications of the stress-strength relationships in industrial and economic systems during the last decades are collected and digested in Kotz, Lumelskii, and Pensky (2003).

The class of life-time distributions (in particular, exponential and gamma) is considered by Nadarajah (2003). Estimation of R from exponential case with common location parameter (Baklizi & El-Masri, 2004), Burr-III (Mokhlis, 2005), beta (Nadarajah, 2005a), gamma (Nadarajah, 2005b), bivariate exponential (Nadarajah & Kotz, 2006), and Weibull (Kundu & Gupta, 2006) distributions were also studied. Inferences on reliability in two-parameter exponential stress-strength model (Krishnamoorthy, Mukherjee, & Guo, 2007) and ML estimation of system reliability for Gompertz distribution (Saraçoglu & Kaya, 2007) were considered. Kakade, Shirke, and Kundu (2008) studied the exponentiated Gumbel case. Baklizi (2008a; 2008b) studied some inference problems of estimating R based on record values. Kundu and Raqab (2009) considered the estimation of R when X and Y are independent and both having three parameter Weibull distribution with common shape and location parameters, but different scale parameters.

Gupta and Peng (2009) studied the estimation of R in the context of proportional odds ratio model. Wang and Shi (2010) obtained the empirical Bayes

inference for the Burr type XII distribution based on records. Lomax (1954) used this distribution in the analysis of business failure data. Balkema and De Haan (1974) showed that this distribution arises as a limit distribution of residual lifetime at great age. The Lomax distribution includes increasing and decreasing hazard rates as well, and was shown to be useful for modeling and analyzing the life time data in medical and biological sciences and engineering, etc.

Many statistical methods have been developed for this distribution. For a review of the Lomax distribution, see Habibullah and Ahsanullah (2000), Upadhyay and Peshwani (2003), and Abd Ellah (2003; 2006) and references therein. A great deal of research was done on estimating the parameters of a Lomax distribution using both classical and Bayesian techniques. The form of the probability density function (pdf) and cumulative distribution function (cdf) with the scale parameter λ and the shape parameter α of the Lomax distribution, denoted by $\text{Lomax}(\lambda, \alpha)$, are given, respectively, by

$$f(x) = \alpha \lambda^\alpha (x + \lambda)^{-(\alpha+1)}, \quad x > 0, \quad \alpha, \lambda > 0 \quad (1)$$

$$F(x) = 1 - \lambda^\alpha (x + \lambda)^{-\alpha}, \quad x > 0, \quad \alpha, \lambda > 0 \quad (2)$$

Record data arise in a wide variety of practical situations. Examples include industrial stress testing, meteorological analysis, hydrology, seismology, sporting, athletic events, and oil and mining surveys. Specifically, Let $\{X_j, j \geq 1\}$ be a sequence of independent identically continuous random variables. An observation X_j will be called an upper record value if its value exceeds that of all previous observations. That is, X_j is an upper record if $X_j > X_i$ for every $i < j$. An analogous definition can be given for lower record values. Record values can be viewed as order statistics from a sample whose size is determined by the values and the order of occurrence of the observations.

Maximum Likelihood Estimator of R

Suppose that X is the strength of a component which is subject to stress Y . The system fails if and only if, at any time, the applied stress is greater than strength. Let X be a random variable whose pdf is a Lomax distribution with parameters λ and α , denoted by $\text{Lomax}(\lambda, \alpha)$, and Y is another Lomax distribution random variable with parameters λ and β , denoted by $\text{Lomax}(\lambda, \beta)$, where X and Y are independent. Thus,

$$R = P[Y < X] = \int_0^{\infty} P(Y < X | X = x) P(X = x) dx$$

$$= \frac{\alpha}{\beta + \alpha} \quad (3)$$

The interest is in estimating R when the data available on both X and Y are in the form of upper record values. To compute the MLE of R , compute the MLE of α and β . Suppose $\mathbf{x} = x_{U(1)}, x_{U(2)}, \dots, x_{U(n)}$ is the first upper record values of size n from $\text{Lomax}(\lambda, \alpha)$, and $\mathbf{y} = y_{U(1)}, y_{U(2)}, \dots, y_{U(m)}$ is an independent set of the first upper record values of size m from $\text{Lomax}(\lambda, \beta)$. The likelihood functions for both observed records $\mathbf{x}$ and $\mathbf{y}$ are given, respectively, by

$$L_1(\lambda, \alpha | \mathbf{x}) = f(x_{U(n)}) \prod_{i=1}^{n-1} \frac{f(x_{U(i)})}{1 - F(x_{U(i)})} \quad (4)$$

and

$$L_2(\lambda, \beta | \mathbf{y}) = g(y_{U(m)}) \prod_{j=1}^{m-1} \frac{g(y_{U(j)})}{1 - G(y_{U(j)})}, \quad (5)$$

where f and F are the pdf and cdf of $X \sim \text{Lomax}(\lambda, \alpha)$, respectively, and g and G are the pdf and cdf of $Y \sim \text{Lomax}(\lambda, \beta)$, respectively (Arnold, Balakrishnan, & Nagaraja, 1998). Substituting f , F , g and G in the likelihood functions,

$$L_1(\lambda, \alpha | \mathbf{x}) = \alpha^n \lambda^\alpha \eta_1(\mathbf{x}, \lambda) \exp \left[-\alpha \ln(x_{U(n)} + \lambda) \right] \quad (6)$$

$$L_2(\lambda, \beta | \mathbf{y}) = \beta^m \lambda^\beta \eta_2(\mathbf{y}, \lambda) \exp \left[-\beta \ln(y_{U(m)} + \lambda) \right], \quad (7)$$

where

$$\eta_1(\mathbf{x}, \lambda) = \prod_{i=1}^n (x_{U(i)} + \lambda)^{-1}, \quad \eta_2(\mathbf{y}, \lambda) = \prod_{j=1}^m (y_{U(j)} + \lambda)^{-1}. \quad (8)$$

BAYESIAN ESTIMATION OF $P[Y < X]$ BASED ON RECORD VALUES

Therefore, the joint Log-likelihood function of the observed records $\mathbf{x}$ and $\mathbf{y}$ under Lomax distribution is

$$\begin{aligned} \ell(\lambda, \alpha, \beta \mid \text{data}) = & n \ln \alpha + m \ln \beta + (\alpha + \beta) \ln \lambda - \alpha \ln(x_{U(n)} + \lambda) \\ & - \beta \ln(y_{U(m)} + \lambda) - \sum_{i=1}^n \ln(x_{U(i)} + \lambda) - \sum_{j=1}^m \ln(y_{U(j)} + \lambda) \end{aligned} \quad (9)$$

The MLEs of λ , α , and β , say $\hat{\lambda}$, $\hat{\alpha}$, and $\hat{\beta}$, respectively, can be obtained as the solution of

$$\frac{\hat{\alpha} + \hat{\beta}}{\hat{\lambda}} - \frac{\hat{\alpha}}{x_{U(n)} + \hat{\lambda}} - \frac{\hat{\beta}}{y_{U(m)} + \hat{\lambda}} - \sum_{i=1}^n \frac{1}{x_{U(i)} + \hat{\lambda}} - \sum_{j=1}^m \frac{1}{y_{U(j)} + \hat{\lambda}} = 0 \quad (10)$$

$$\frac{n}{\hat{\alpha}} - \ln(x_{U(n)} + \hat{\lambda}) + \ln \hat{\lambda} = 0 \quad (11)$$

$$\frac{m}{\hat{\beta}} - \ln(y_{U(m)} + \hat{\lambda}) + \ln \hat{\lambda} = 0 . \quad (12)$$

From (11) and (12),

$$\hat{\alpha} = n \left[\ln \left(\frac{x_{U(n)}}{\hat{\lambda}} \right) + 1 \right]^{-1} \quad (13)$$

$$\hat{\beta} = m \left[\ln \left(\frac{y_{U(m)}}{\hat{\lambda}} \right) + 1 \right]^{-1} , \quad (14)$$

and $\hat{\lambda}$ can be obtained as the solution of the following non-linear equation:

$$\begin{aligned}
 f(\lambda) = & \frac{n}{\lambda} \left[\ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right]^{-1} + \frac{m}{\lambda} \left[\ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right]^{-1} \\
 & - \frac{n}{x_{U(n)} + \lambda} \left[\ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right]^{-1} - \frac{m}{y_{U(m)} + \lambda} \left[\ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right]^{-1} \\
 & - \sum_{i=1}^n \frac{1}{x_{U(i)} + \lambda} - \sum_{j=1}^m \frac{1}{y_{U(j)} + \lambda}
 \end{aligned} \quad (15)$$

Therefore, $\hat{\lambda}$ can be obtained as the solution of the non-linear equation of the form

$$h(\lambda) = \lambda \quad (16)$$

where

$$\begin{aligned}
 h(\lambda) = & \left(n \left[\ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right]^{-1} + m \left[\ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right]^{-1} \right) \left[\frac{n}{x_{U(n)} + \lambda} \left[\ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right]^{-1} \right. \\
 & \left. + \frac{m}{y_{U(m)} + \lambda} \left[\ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right]^{-1} + \sum_{i=1}^n \frac{1}{x_{U(i)} + \lambda} + \sum_{j=1}^m \frac{1}{y_{U(j)} + \lambda} \right]^{-1}
 \end{aligned}$$

Since $\hat{\lambda}$ is a fixed point solution of non-linear equation (15) and, therefore, it can be obtained by using a simple iterative scheme as follows:

$$h(\lambda_j) = \lambda_{j+1}, \quad (17)$$

where λ_j is the j^{th} iterate of $\hat{\lambda}$. The iteration procedure should be stopped when $|\lambda_{j+1} - \lambda_j|$ is sufficiently small. Therefore, the MLE of R becomes

$$\hat{R} = \frac{\hat{\alpha}}{\hat{\beta} + \hat{\alpha}} \quad (18)$$

Bootstrap Confidence Intervals

Confidence intervals are proposed based on the parametric bootstrap methods (i) percentile bootstrap method (Boot-p) based on the idea of Efron (1982) and (ii) bootstrap-t method (Boot-t) based on the idea of Hall (1988). The algorithms for estimating the confidence intervals of R using both methods are illustrated as follows:

Percentile Bootstrap Method

1. From the original two samples of upper records $\{x_{U(1)}, x_{U(2)}, \dots, x_{U(n)}\}$ and $\{y_{U(1)}, y_{U(2)}, \dots, y_{U(m)}\}$, compute ML estimates $\hat{\lambda}$, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{R}$
2. Using $\hat{\lambda}$ and $\hat{\alpha}$, generate a bootstrap upper record sample $\{x_{U(1)}^*, x_{U(2)}^*, \dots, x_{U(n)}^*\}$ and, similarly, using $\hat{\lambda}$ and $\hat{\beta}$, generate a bootstrap upper record sample $\{y_{U(1)}^*, y_{U(2)}^*, \dots, y_{U(m)}^*\}$. Based on these data, we compute the bootstrap estimates, say $\hat{\lambda}^*$, $\hat{\alpha}^*$, $\hat{\beta}^*$, and $\hat{R}^*$
3. Repeat step 2 N_{boot} times
4. Let $V(x) = P(\hat{R}^* \leq x)$ be the cdf of $\hat{R}^*$. Define $\hat{R}_{boot} = V^{-1}(x)$ for a given x . The approximate $100(1 - \gamma)\%$ confidence interval of R is given by

$$\left(\hat{R}_{Boot-p} \left(\frac{\gamma}{2} \right), \hat{R}_{Boot-p} \left(1 - \frac{\gamma}{2} \right) \right). \quad (19)$$

Bootstrap-t Method

1. From the original two samples of upper records $\{x_{U(1)}, x_{U(2)}, \dots, x_{U(n)}\}$ and $\{y_{U(1)}, y_{U(2)}, \dots, y_{U(m)}\}$, compute ML estimates $\hat{\lambda}$, $\hat{\alpha}$, $\hat{\beta}$, and $\hat{R}$
2. Using $\hat{\lambda}$ and $\hat{\alpha}$, generate a bootstrap upper record sample $\{x_{U(1)}^*, x_{U(2)}^*, \dots, x_{U(n)}^*\}$ and, similarly, using $\hat{\lambda}$ and $\hat{\beta}$, generate a bootstrap upper record sample $\{y_{U(1)}^*, y_{U(2)}^*, \dots, y_{U(m)}^*\}$. Based on these data, we compute the bootstrap estimates, say $\hat{\lambda}^*$, $\hat{\alpha}^*$, $\hat{\beta}^*$, and compute the bootstrap estimate of R using (18), $\hat{R}^*$, and following statistic:

$$T^* = \frac{\sqrt{n}(\hat{R}^* - \hat{R})}{\sqrt{\text{Var}(\hat{R}^*)}},$$

where $\text{Var}(\hat{R}^*)$ is obtained using the Delta method (Greene, 2000)

3. Repeat step 2 N boot times
4. For the T^* values obtained in step 2, determine the upper and lower bounds of the $100(1 - \gamma)\%$ confidence interval of R as follows: let $H(x) = P(T^* \leq x)$ be the cdf of T^* . For a given x , define

$$\hat{R}_{\text{Boot-t}}(x) = \hat{R} + n^{-1/2} \sqrt{\text{Var}(\hat{R})} H^{-1}(x).$$

Here also, $\text{Var}(\hat{R})$ can be computed as same as computing the $\text{Var}(\hat{R}^*)$.

The approximate $100(1 - \gamma)\%$ confidence interval of R is given by

$$\left(\hat{R}_{\text{Boot-t}}\left(\frac{\gamma}{2}\right), \hat{R}_{\text{Boot-t}}\left(1 - \frac{\gamma}{2}\right) \right) \quad (20)$$

Bayes Estimation of R Using MCMC

The advantage of MCMC is that it not only gives a point estimate of the parameter, but also gives an interval estimation based on the final simulated empirical distribution. MCMC is essentially an iterative sampling algorithm, drawing values from the posterior distributions of the parameter in the concerned model. Consider the MCMC method to generate samples from the posterior distributions and then compute the Bayes estimates of R under upper record values from Lomax distribution. A wide variety of MCMC schemes are available. An important subclass of MCMC methods are Gibbs sampling and more general Metropolis-Hastings (M-H) algorithm (Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953; Hastings, 1970). For more details about MCMC and the related methodologies, one can refer to Gentle (1998), Chen, Shao, and Ibrahim (2000), and Robert and Casella (2004).

Now, obtain the Bayes estimation of R under assumption that the parameters (λ, α, β) are independent random variables. The Bayes estimate of R under the squared error loss and the corresponding credible interval by the Gibbs sampling

technique are considering. It is assumed that (λ, α, β) have independent gamma priors with the pdfs

$$\pi_1(\lambda | a_1, b_1) = \begin{cases} \frac{b_1^{a_1}}{\Gamma(a_1)} \lambda^{a_1-1} \exp(-b_1 \lambda), & \lambda > 0 \\ 0, & \lambda \leq 0 \end{cases} \quad (21)$$

$$\pi_2(\alpha | a_2, b_2) = \begin{cases} \frac{b_2^{a_2}}{\Gamma(a_2)} \alpha^{a_2-1} \exp(-b_2 \alpha), & \alpha > 0 \\ 0, & \alpha \leq 0 \end{cases} \quad (22)$$

$$\pi_3(\beta | a_3, b_3) = \begin{cases} \frac{b_3^{a_3}}{\Gamma(a_3)} \beta^{a_3-1} \exp(-b_3 \beta), & \beta > 0 \\ 0, & \beta \leq 0 \end{cases} \quad (23)$$

where, a_1, b_1, a_2, b_2, a_3 , and b_3 are chosen to reflect prior knowledge about λ, α , and β . The expression for the posterior can be obtained up to proportionality by multiplying the likelihood with the prior and this can be written as

$$\begin{aligned} \pi^*(\lambda, \alpha, \beta | \text{data}) &\propto \lambda^{(\alpha+\beta+a_1-1)} \alpha^{(n+a_2-1)} \beta^{(m+a_3-1)} \eta_1(\mathbf{x}, \lambda) \eta_2(\mathbf{y}, \lambda) \exp(-b_1 \lambda) \\ &\times \exp\left[-\alpha \left(b_2 + \ln(x_{U(n)} + \lambda)\right)\right] \\ &\times \exp\left[-\beta \left(b_3 + \ln(y_{U(m)} + \lambda)\right)\right] \end{aligned} \quad (24)$$

where $\eta_1(\mathbf{x}, \lambda)$, $\eta_2(\mathbf{y}, \lambda)$ are given in (8) and

$$\begin{aligned} \pi_1^*(\lambda | \alpha, \beta, \text{data}) &\propto \lambda^{(\alpha+\beta+a_1-1)} (x_{U(n)} + \lambda)^{-\alpha} (y_{U(m)} + \lambda)^{-\beta} \\ &\times \eta_1(\mathbf{x}, \lambda) \eta_2(\mathbf{y}, \lambda) \exp(-b_1 \lambda) \end{aligned} \quad (25)$$

Similarly, the full posterior conditional distribution for α and β are given by

$$\pi_2^*(\alpha | \lambda, \text{data}) \sim \text{Gamma} \left(n + a_2, b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right) \quad (26)$$

$$\pi_3^*(\beta | \lambda, \text{data}) \sim \text{Gamma} \left(m + a_3, b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right). \quad (27)$$

It can be seen that (26) is a gamma density with shape parameters $(n + a_2)$ and $\left(b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right)$ and (27) is a gamma density with shape parameters $(m + a_2)$ and $\left(b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right)$. Therefore, samples of α and β can be easily generated using any gamma generating routine. However, in our case, the conditional posterior distribution of λ equation (25) cannot be reduced analytically to well-known distributions and, therefore, it is not possible to sample directly by standard methods. However, the plot of it shows that it is similar to a normal distribution. To generate random numbers from this distribution, use the Metropolis-Hastings method with normal proposal distribution.

Therefore, the algorithm of Gibbs sampling is as follows:

- Start with an $(\lambda^{(0)} = \hat{\lambda})$ and set $t = 1$
- Generate $\alpha^{(t)}$ from $\text{Gamma} \left(n + a_2, b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right)$
- Generate $\beta^{(t)}$ from $\text{Gamma} \left(m + a_3, b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right)$
- Using Metropolis-Hastings (see [Metropolis et al., 1953](#)), generate $\lambda^{(t)}$ from (25) with the $N(\lambda^{(t-1)}, \sigma^2)$ proposal distribution, where σ^2 is variance-covariance matrix
- Compute $\lambda^{(t)}$, $\alpha^{(t)}$, and $\beta^{(t)}$. Then compute $R^{(t)} = \frac{\alpha^{(t)}}{\beta^{(t)} + \alpha^{(t)}}$
- Set $t = t + 1$
- Repeat steps 2-5 N times
- We obtain the Bayes MCMC point estimate of R as

$$E(R | \text{data}) = \frac{1}{N-M} \sum_{i=M+1}^N R^{(i)}, \quad (28)$$

where M is the burn-in period (that is, a number of iterations before the stationary distribution is achieved) and posterior variance of R becomes

$$\hat{V}(R | \text{data}) = \frac{1}{N-M} \sum_{i=M+1}^N \left(R^{(i)} - \hat{E}(R | \text{data}) \right)^2 \quad (29)$$

- To compute the credible intervals of R , usually, take the quintiles of the sample as the end points of the interval. Order $R^{(M+1)}, R^{(M+2)}, \dots, R^{(N)}$ as $R^{(1)}, R^{(2)}, \dots, R^{(N-M)}$. Then the $100(1-\gamma)\%$ symmetric credible interval is

$$\left(R_{\left(\frac{\gamma}{2}(N-M)\right)}, R_{\left(\left(1-\frac{\gamma}{2}\right)(N-M)\right)} \right) \quad (30)$$

Estimation of R if λ is Known

Consider the estimation of R and the corresponding highest posterior density (HPD) intervals when λ is known. Assume $x_{U(1)}, x_{U(2)}, \dots, x_{U(n)}$ is the first upper record values observed from $\text{Lomax}(\lambda, \alpha)$, and $y_{U(1)}, y_{U(2)}, \dots, y_{U(m)}$ is the first upper record values observed from $\text{Lomax}(\lambda, \beta)$. Based on these samples, we can estimate R . Recently works on interval estimation of R were discussed in Rezaeia, Tahmasbib, and Mahmoodi (2010), Baklizi (2008a; 2008b), and Shoukri, Chaudhary, and Al-Halees (2005). First, consider the MLE of R and its distributional properties.

MLE of R

The MLE of R , $\hat{R}$, will be

$$\hat{R} = \frac{\hat{\alpha}}{\hat{\beta} + \hat{\alpha}}, \quad (31)$$

where

$$\hat{\alpha} = n \left[\ln \left(\frac{x_{U(n)}}{\hat{\lambda}} + 1 \right) \right]^{-1}, \quad \hat{\beta} = m \left[\ln \left(\frac{y_{U(m)}}{\hat{\lambda}} + 1 \right) \right]^{-1}. \quad (32)$$

Therefore,

$$\hat{R} = \frac{n \ln \left(\frac{y_{U(m)}}{\hat{\lambda}} + 1 \right)}{m \ln \left(\frac{x_{U(n)}}{\hat{\lambda}} + 1 \right) + n \ln \left(\frac{y_{U(m)}}{\hat{\lambda}} + 1 \right)}. \quad (33)$$

To study the confidence interval of R , consider the distribution of $\hat{R}$ as well as the distributions of $\hat{\alpha}$ and $\hat{\beta}$. Consider first $\hat{\alpha} = n \left[\ln \left(\frac{x_{U(n)}}{\hat{\lambda}} + 1 \right) \right]^{-1}$. Arnold et al. (1998) obtained the pdf of R_n as follows:

$$f_{R_n}(r_n) = f(r_n) \left[-\ln(1 - F(r_n)) \right]^{n-1} / (n-1)! \quad (34)$$

under Lomax(λ, α)

$$f_{R_n}(r_n) = \frac{\alpha \lambda^\alpha}{\Gamma(n)} (r_n + \lambda)^{-(\alpha+1)} \left[\alpha \ln \left(\frac{r_n}{\lambda} + 1 \right) \right]^{n-1}, \quad r_n > 0. \quad (35)$$

Consequently, the pdf of $Z_1 = \hat{\alpha}$ as defined in (32) is given by

$$f_{Z_1}(Z_1) = \frac{(n\alpha)^n}{\Gamma(n) Z_1^{n+1}} \exp \left(-\frac{n\alpha}{Z_1} \right), \quad Z_1 > 0. \quad (36)$$

This is the inverted gamma distribution. Similarly, the pdf of $Z_2 = \hat{\beta}$ as defined in (32) is given by

$$f_{Z_2}(Z_2) = \frac{(m\beta)^m}{\Gamma(m) Z_2^{m+1}} \exp \left(-\frac{m\beta}{Z_2} \right), \quad Z_2 > 0. \quad (37)$$

Find the pdf of

$$\hat{R} = \frac{\hat{\alpha}}{\hat{\beta} + \hat{\alpha}} = \frac{Z_1}{Z_1 + Z_2} = \frac{1}{1 + Z_2/Z_1}.$$

Consider Z_2/Z_1 . By the properties of the inverted gamma distribution and its relation with the gamma distribution, $n\alpha/Z_1 \sim \text{Gamma}(n, 1)$ and $m\beta/Z_2 \sim \text{Gamma}(m, 1)$. Hence $2n\alpha/Z_1 \sim \chi_{2n}^2$ and $2m\beta/Z_2 \sim \chi_{2m}^2$. By the independence of two random quantities,

$$\frac{2n\alpha/2nZ_1}{2m\beta/2mZ_2} \sim F_{2m, 2n},$$

and thus $\frac{Z_2}{Z_1} = \frac{\beta}{\alpha} F_{2m, 2n}$, a scaled F distribution. It follows that the distribution of

$\hat{R}$ is that of $\frac{1}{1 + \frac{\beta}{\alpha} F_{2m, 2n}}$. Because $\frac{1}{\hat{R}} = 1 + \left(\frac{1}{R} - 1\right) F(2m, 2n)$, then

$$R \sim \frac{F(2m, 2n)}{F(2m, 2n) + \left(\frac{1}{\hat{R}} - 1\right)}.$$

The $100(1 - \gamma)\%$ confidence interval of R can be obtained as

$$\left[\frac{F_{\gamma/2}(2m, 2n)}{F_{\gamma/2}(2m, 2n) + \left(\frac{1}{\hat{R}} - 1\right)}, \frac{F_{1-\gamma/2}(2m, 2n)}{F_{1-\gamma/2}(2m, 2n) + \left(\frac{1}{\hat{R}} - 1\right)} \right]. \quad (38)$$

Bayes Estimation of R

Obtain the Bayes estimation of R under assumption that the shape parameters α and β are random variables. It is assumed that α and β have independent gamma priors (22) and (23), respectively, with the parameters $\alpha \sim \text{Gamma}(a_2, b_2)$ and

$\beta \sim \text{Gamma}(a_3, b_3)$. The posterior pdfs of α and β are given by (26) and (27), respectively, because priors α and β are independent. Using standard transformation techniques, and after some manipulations, the posterior pdf of $R = \alpha/\beta + \alpha$ is found to be

$$f_R(r) = C \frac{r^{n+a_2-1} (1-r)^{n+a_3-1}}{\left[r \left(b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right) + (1-r) \left(b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right) \right]^{n+m+a_2+a_3}} \quad (39)$$

if $0 < r < 1$, and 0 otherwise, where

$$C = \frac{\Gamma(n+m+a_2+a_3)}{\Gamma(n+a_2)\Gamma(m+a_3)} \left[b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right) \right]^{n+a_2} \left[b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right) \right]^{m+a_3}. \quad (40)$$

There is no explicit expression for the posterior mean or median of (39). However, the posterior mode can be easily obtained as

$$\frac{d}{dr} f_R(r) = \frac{r^{A_1-1} (1-r)^{A_2-1} \left[2r^2 (B_1 - B_2) + r(2B_2 - 2B_1 - A_2 B_1 + A_1 B_2) + A_1 B_2 \right]}{\left[B_2 (1-r) + B_1 r \right]^{(A_1+A_2+3)}},$$

where $B_1 = b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right)$, $B_2 = b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right)$, $A_1 = n+a_2-1$, and

$A_2 = m+a_3-1$. Note $r \in (0, 1)$, $(d/dr)f_R(r) = 0$ has only two roots. Using the fact that $\lim_{r \rightarrow 0^+} (d/dr)f_R(r) > 0$ and $\lim_{r \rightarrow 1^-} (d/dr)f_R(r) < 0$, it follows that the density function $f_R(r)$ has a unique mode. The posterior mode can be obtained as the unique root of which lies between 0 and 1 of the following quadratic equation:

$$2r^2 (B_2 - B_1) + r(2B_1 - 2B_2 + A_2 B_1 - A_1 B_2) - A_1 B_2 = 0. \quad (41)$$

Consider the following loss function:

$$L(a, b) = \begin{cases} 1, & |a - b| > \varepsilon \\ 0, & |a - b| \leq \varepsilon \end{cases} \quad (42)$$

It is known that the Bayes estimate with respect to the above loss function (42) is the midpoint of the ‘modal interval’ of length 2ε of the posterior distribution. Therefore, the posterior mode is an approximate Bayes estimator of R with respect to the above loss function when the constant ε is small.

The Bayes estimate of R under squared error loss cannot be computed analytically. Alternatively, using the approximate method of Lindley (1980), it can be easily seen that the approximate Bayes estimate of R , say $\tilde{R}_{\text{Bayes}}$, relative to squared error loss function is

$$\tilde{R}_{\text{Bayes}} = \tilde{R} \left[1 + \frac{\tilde{\alpha} \tilde{R}^2}{\tilde{\beta}^2 (n + a_2 - 1)(m + a_3 - 1)} [\tilde{\alpha} (n + a_2 - 1) - \tilde{\beta} (m + a_3 - 1)] \right] \quad (43)$$

where

$$\tilde{\alpha} = \frac{n + a_2 - 1}{b_2 + \ln \left(\frac{x_{U(n)}}{\lambda} + 1 \right)}, \quad \tilde{\beta} = \frac{m + a_3 - 1}{b_3 + \ln \left(\frac{y_{U(m)}}{\lambda} + 1 \right)}, \quad \tilde{R} = \frac{\tilde{\alpha}}{\tilde{\beta} + \tilde{\alpha}}. \quad (44)$$

For comparison purposes, a highest posterior density (HPD) interval of R was computed (Soliman & Al-Aboud, 2008). Due to the unimodality of the posterior distribution (39), the $100(1 - \gamma)\%$ HPD interval $[\omega_L, \omega_U]$ for R is given by the simultaneous solution of the nonlinear equations

$$\int_{\omega_L}^{\omega_U} f_R(r | \text{dat}) dr = 1 - \gamma, \quad f_R(\omega_L | \text{dat}) = f_R(\omega_U | \text{dat}) \quad (45)$$

A Newton-Raphson iteration can be invoked to solve the equations in (45) and thereby the HPD interval is obtained.

Illustrative Example Using Simulated Data

Six upper record values were simulated from Lomax(1, 2.1) and six upper record values from Lomax(1, 2.5), with $R_{\text{Exact}} = 0.4565$. The data were truncated after four

decimal places: the x upper record values are $\{1.0638, 1.4488, 7.2166, 7.8652, 11.6919, 34.5528\}$ and the corresponding y upper record values are $\{0.2355, 1.0058, 1.5503, 2.0698, 12.8867, 13.0820\}$.

Case (1)

λ is unknown: Based on the above data, plot the profile log-likelihood function of λ in Figure 2. It is an upside down function and it has a unique maximum. Obtain the MLE of λ using the iterative procedure (16). Using the stopping criterion that the iteration stops whenever two consecutive values are less than 10^{-6} , the iteration stops after 14 steps and it provides the MLE of $\hat{\lambda} = 1.5232$. Using (13) and (14), obtain the MLEs of $\hat{\alpha} = 1.8958$ and $\hat{\beta} = 2.6542$, and hence $\hat{R} = 0.4167$, from (18). The 95% confidence, credible intervals, and corresponding length are reported in Table 1 using exact confidence interval, parametric percentile bootstrap methods, and MCMC technique.

Case (2)

Estimate the parameters assuming λ is known to be 1. Obtain the MLEs of α and β as 1.7079 and 2.4080, respectively. Therefore, the MLE of R becomes $\hat{R} = 0.4150$. The corresponding 95% confidence credible intervals and corresponding length are also reported in Table 1 using MLE, parametric percentile bootstrap methods, and MCMC technique. The posterior probability density function (39) of R for the given data set is plotted in Figure 2. The simulation number of R and Histogram of R generated by MCMC method are plotted in Figure 1.

Table 1. Two-sided 95 % confidence and credible intervals of R when $\lambda = 1$, $\alpha = 2.1$, and $\beta = 2.5$ with prior 0

Methods	λ is unknown			λ is known		
	$\hat{R}$	95% CI	Length	$\hat{R}$	95% CI	Length
MLE	0.4167	[0.1789, 0.7007]	0.5217	0.4150	[0.1779, 0.6992]	0.5213
Boot-p	0.4131	[0.2150, 0.6246]	0.4096	0.4189	[0.1796, 0.6921]	0.5125
Boot-t	0.4008	[0.0634, 0.6553]	0.5920	0.4036	[0.0271, 0.7306]	0.7035
Bayes	0.4111	[0.1673, 0.6902]	0.5229	0.4192	[0.1883, 0.7135]	0.5252

BAYESIAN ESTIMATION OF $P[Y < X]$ BASED ON RECORD VALUES

Figure 1. Simulation number (left) and histogram (right) of R generated by MCMC method

Figure 2. Profile likelihood of λ (left) and posterior pdf of R (right) for given data set presented above

Simulation Results

Some numerical experiments were performed to evaluate the behavior of the proposed methods for different samples, different parameter values, and different hyper parameters. Consider two cases separately to draw inference on R , namely when (i) common scale parameter λ is unknown and (ii) the common scale parameter λ is known. Consider the different sample sizes (n and m) and different hyper parameters ($a_1, b_1, a_2, b_2, a_3, b_3$). In first case, take $\lambda = 2$, $\alpha = 1.5$, and $\beta = 3$. In the second case, take $\lambda = 2$, $\alpha = 3.2$, and $\beta = 2.1$. Without loss of generality, take $\lambda = 2$ in both cases. All the results are based on 1000 replications.

(i) λ is unknown. Compute the estimate of λ using the iterative algorithm (16) with the initial estimate 2. The iterative process stops when the difference between the two consecutive iterates are less than 10^{-7} . Once λ is obtained, then estimate α

and β using (13) and (14), respectively. Finally, obtain the MLE of R using (18). To find the Bayes MCMC estimates, use the non-informative gamma priors for the three parameters (we call it prior 0). Non-informative prior ($a_1 = b_1 = a_2 = b_2 = a_3 = b_3 = 0$) provides prior distributions which are not proper. Also use informative priors, including prior 1, $a_1 = 2, b_1 = 1, a_2 = 3, b_2 = 2, a_3 = 3$, and $b_3 = 1$, with the values of previous parameters and compute the Bayes estimates and 95% probability intervals based on 10,000 MCMC samples (discard the first 1,000 values as ‘burn-in’). The average Bayes estimates, means squared errors (MSEs), coverage percentages, and average probability interval lengths based on 1000 replications are reported in Table 2.

(ii) λ is known. Obtain the estimates of R by using the ML method and Lindley's approximation approach. Calculate the exact confidence intervals and HPD interval of R , using the same non-informative prior (prior 0) and an informative prior, including (prior 1) to compute the average estimates of R , MSEs, coverage percentages, and average probability interval lengths based on 1,000 replications. The results are reported in Table 3.

Table 2. Simulation results and estimation of the parameters

(n, m)	R_{Exact}	MLE		Bayes using MCMC			
		Mean	MSE	Mean	MSE	Length	Coverage
$\lambda = 2, \alpha = 1.5, \beta = 3$ using prior 0							
(5, 5)	0.3333	0.3184	0.0205	0.3757	0.0115	0.5311	0.9900
(6, 6)		0.3206	0.0203	0.3742	0.0113	0.4863	0.9810
(7, 6)		0.3227	0.0182	0.3696	0.0111	0.4807	0.9750
(7, 7)		0.3459	0.0169	0.3881	0.0110	0.4617	0.9900
(8, 7)		0.3133	0.0154	0.3735	0.0107	0.4431	0.9850
(8, 8)		0.3286	0.0146	0.3775	0.0106	0.4323	0.9950
(9, 8)		0.3326	0.0141	0.3809	0.0104	0.4211	0.9750
(9, 9)		0.3184	0.0140	0.3624	0.0089	0.4004	0.9550
(10, 9)		0.3172	0.0131	0.3682	0.0087	0.3944	0.9550
(10, 10)		0.3224	0.0126	0.3622	0.0082	0.3819	0.9650
$\lambda = 2, \alpha = 1.5, \beta = 3$ using prior 1							
(5, 5)	0.3333	0.3269	0.0202	0.3766	0.0055	0.4399	0.9930
(6, 6)		0.3203	0.0201	0.3698	0.0053	0.4104	0.9950
(7, 6)		0.3246	0.0184	0.3812	0.0052	0.4076	0.9950
(7, 7)		0.3242	0.0151	0.3651	0.0048	0.3895	0.9760
(8, 7)		0.3330	0.0149	0.3780	0.0047	0.3867	0.9950
(8, 8)		0.3378	0.0148	0.3727	0.0045	0.3744	0.9770
(9, 8)		0.3262	0.0146	0.3689	0.0042	0.3653	0.9900
(9, 9)		0.3352	0.0131	0.3663	0.0041	0.3578	0.9660
(10, 9)		0.3335	0.0127	0.3712	0.0039	0.3525	0.9850
(10, 10)		0.3495	0.0124	0.3714	0.0037	0.3443	0.9950

BAYESIAN ESTIMATION OF $P[Y < X]$ BASED ON RECORD VALUES

Table 3. Simulation results and estimation of the parameters

(n, m)	R _{Exact}	MLE				Bayes using Lindely			
		Mean	MSE	Length	Coverage	Mean	MSE	Length	Coverage
$\lambda = 2, \alpha = 3.2, \beta = 2.1$ using prior 0									
(5, 5)	0.6038	0.6407	0.0462	0.4535	0.9550	0.6276	0.0393	0.5851	0.9900
(6, 6)		0.6334	0.0373	0.4361	0.9310	0.6298	0.0286	0.5787	0.9950
(7, 6)		0.5954	0.0369	0.4356	0.9450	0.6356	0.0266	0.5308	0.9700
(7, 7)		0.6288	0.0264	0.4259	0.9600	0.6194	0.0224	0.5294	0.9750
(8, 7)		0.6176	0.0235	0.4162	0.9450	0.6211	0.0219	0.5122	0.9650
(8, 8)		0.6129	0.0213	0.4146	0.9280	0.6241	0.0215	0.4941	0.9800
(9, 8)		0.6099	0.1869	0.4100	0.9330	0.6197	0.0162	0.4677	0.9550
(9, 9)		0.6245	0.0176	0.3930	0.9010	0.6407	0.0149	0.4512	0.9390
(10, 9)		0.6021	0.0170	0.3922	0.8990	0.6378	0.0078	0.4189	0.9400
(10, 10)		0.6059	0.0165	0.3820	0.9090	0.6395	0.0052	0.3972	0.9600
$\lambda = 2, \alpha = 3.2, \beta = 2.1$ using prior 0									
(5, 5)	0.6038	0.5997	0.0454	0.4732	0.9610	0.6084	0.0251	0.5933	0.9450
(6, 6)		0.5964	0.0397	0.4469	0.9800	0.5832	0.0245	0.5532	0.9230
(7, 6)		0.6156	0.0332	0.4348	0.9120	0.5985	0.0214	0.5159	0.9200
(7, 7)		0.6001	0.0327	0.4221	0.9330	0.5979	0.0207	0.5144	0.9450
(8, 7)		0.6077	0.0260	0.4201	0.9350	0.6193	0.0168	0.5083	0.9450
(8, 8)		0.6065	0.0232	0.4131	0.9050	0.6115	0.0162	0.4899	0.9600
(9, 8)		0.6236	0.0191	0.4022	0.9010	0.6261	0.0136	0.4570	0.9850
(9, 9)		0.6201	0.0157	0.3986	0.9200	0.6283	0.0130	0.4047	0.9640
(10, 9)		0.6029	0.0111	0.3930	0.9050	0.6296	0.0089	0.3995	0.9540
(10, 10)		0.6016	0.0093	0.3822	0.9150	0.6194	0.0051	0.3755	0.9600

Conclusion

The problem of estimating $R = P[Y < X]$ for the Lomax distributions was addressed, and classical and MCMC Bayesian analysis for R were developed when both samples on X and Y are in the form of upper record values, observed from the Lomax distribution with different one shape parameter. The general case when all the parameters are unknown was considered, and when the common scale parameter was known. In the first case, the MCMC method provided an alternative method for parameters estimation of the Lomax distribution and, also, for obtaining both point and interval estimators of the stress-strength reliability model R . It is more flexible when compared with the traditional methods, such as MLE, based on the set of upper record values. It is hoped this investigation will be useful for researchers dealing with the kind of data considered.

Observe the following from the results:

- When the common scale parameter λ is unknown, it is observed that the Bayes estimator using MCMC technique works quite well. The MCMC sample were used to construct confidence intervals and that also works quite well. When the common scale parameter λ is known, the maximum likelihood estimator and Bayes estimators were proposed based on the approximate method of Lindley. The confidence interval based on the exact distribution of the MLE works quite very well. Also, a HPD interval was recommended
- Tables 2 and 3 show that, when $m = n$ and m, n increase, then MSEs and average confidence interval lengths, credible interval lengths of the MLEs, and Bayes estimators decrease, and that the coverage percentages are reached to the nominal level in most cases
- From Tables 2 and 3, it is clear that the Bayes estimators based on informative priors (prior 1) perform much better than the Bayes estimators based on non-informative priors (prior 0) or MLEs in terms of biases, MSEs, and lengths of credible intervals.

References

- Abd Ellah, A. H. (2003). Bayesian one sample prediction bounds for the Lomax distribution. *Indian Journal Pure and Applied Mathematics*, 34(1), 101-109.
- Abd Ellah, A. H. (2006). Comparison of estimates using record statistics from Lomax model: Bayesian and non Bayesian approaches. *Journal of Statistical Research of Iran*, 3(2), 139-158. Retrieved from <http://en.journals.sid.ir/ViewPaper.aspx?ID=134210>
- Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1998). *Records*. New York, NY: Wiley.
- Baklizi, A., & El-Masri A. Q. (2004). Shrinkage estimation of $P(X < Y)$ in the exponential case with common location parameter. *Metrika*, 59(2), 163-171. doi: 10.1007/s001840300277
- Baklizi, A. (2008a). Estimation of $\Pr(X < Y)$ using record values in the one and two parameter exponential distributions. *Communications in Statistics – Theory and Methods*, 37(5), 692-698. doi: 10.1080/03610920701501921
- Baklizi, A. (2008b). Likelihood and Bayesian estimation of $\Pr(X < Y)$ using lower record values from the generalized exponential distribution. *Computational Statistics & Data Analysis*, 52(7), 3468-3473. doi: 10.1016/j.csda.2007.11.002

- Chen, M. H., Shao, Q. M., & Ibrahim, J. G. (2000). *Monte Carlo methods in Bayesian computation*. New York, NY: Springer-Verlag.
- Efron, B. (1982). The jackknife, the bootstrap and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics*, CB38. doi: [10.1137/1.9781611970319.fm](https://doi.org/10.1137/1.9781611970319.fm)
- Gentle, J. E. (1998). *Random number generation and Monte Carlo methods*. New York, NY: Springer.
- Gupta, R. C., & Peng, C. (2009). Estimating reliability in proportional odds ratio models. *Computational Statistics & Data Analysis*, 53(4), 1495-1510. doi: [10.1016/j.csda.2008.10.014](https://doi.org/10.1016/j.csda.2008.10.014)
- Habibullah, M., & Ahsanullah, M. (2000). Estimation of parameters of a Pareto distribution by generalized order statistics. *Communication in Statistics – Theory and Methods*, 29(7), 1597-1609. doi: [10.1080/03610920008832567](https://doi.org/10.1080/03610920008832567)
- Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, 16(3), 927-953. Available from <http://www.jstor.org/stable/2241604>
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1), 97-109. doi: [10.1093/biomet/57.1.97](https://doi.org/10.1093/biomet/57.1.97)
- Kakade, C. S., Shirke, D. T., & Kundu, D. (2008). Inference for $P(Y < X)$ in exponentiated Gumbel distribution, *Journal of Statistics and Applications*, 3(1-2): 121-133.
- Kotz, S., Lumelskii, Y., & Pensky, M. (2003). *The stress-strength model and its generalizations: Theory and applications: $P(X < Y)$* . Singapore: World Scientific.
- Krishnamoorthy, K., Mukherjee, S., & Guo, H. (2007). Inference on reliability in two-parameter exponential stress-strength model. *Metrika*, 65(3), 261-273. doi: [10.1007/s00184-006-0074-7](https://doi.org/10.1007/s00184-006-0074-7)
- Kundu, D., & Gupta, R. D. (2005). Estimation of $P[Y < X]$ for generalized exponential distribution. *Metrika*, 61(3), 291-308. doi: [10.1007/s001840400345](https://doi.org/10.1007/s001840400345)
- Kundu, D., & Gupta, R. D. (2006). Estimation of $P[Y < X]$ for Weibull distributions. *IEEE Transactions on Reliability*, 55(2), 270-280. doi: [10.1109/TR.2006.874918](https://doi.org/10.1109/TR.2006.874918)
- Kundu, D., & Raqab, M. Z. (2009). Estimation of $R=P(Y < X)$ for three-parameter Weibull distribution. *Statistics & Probability Letters*, 79(17), 1839-1846. doi: [10.1016/j.spl.2009.05.026](https://doi.org/10.1016/j.spl.2009.05.026)

- Lindley, D. V. (1980). Approximate Bayesian methods. *Trabajos de Estadística Y de Investigacion Operativa*, 31(1), 223-237. doi: [10.1007/BF02888353](https://doi.org/10.1007/BF02888353)
- Lomax, K. S. (1954) Business failure: Another example of the analysis of the failure data. *Journal of the American Statistical Association*, 49(268), 847-852. doi: [10.1080/01621459.1954.10501239](https://doi.org/10.1080/01621459.1954.10501239)
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6), 1087-1091. doi: [10.1063/1.1699114](https://doi.org/10.1063/1.1699114)
- Mokhlis, N. A. (2005). Reliability of a stress-strength model with Burr type III distributions. *Communications in Statistics – Theory and Methods*, 34(7), 1643-1657. doi: [10.1081/STA-200063183](https://doi.org/10.1081/STA-200063183)
- Nadarajah, S. (2003). Reliability for lifetime distributions. *Mathematical and Computer Modelling*, 37(7-8), 683-688. doi: [10.1016/S0895-7177\(03\)00074-8](https://doi.org/10.1016/S0895-7177(03)00074-8)
- Nadarajah, S. (2004a). Reliability for logistic distributions. *Elektronnoe Modelirovanie*, 26(3), 65-82.
- Nadarajah, S. (2004b). Reliability for Laplace distributions. *Mathematical Problems in Engineering*, 2004(2), 169-183. doi: [10.1155/S1024123X0431104X](https://doi.org/10.1155/S1024123X0431104X)
- Nadarajah, S. (2005a). Reliability for some bivariate beta distributions. *Mathematical Problems in Engineering*, 2005(1), 101-111. doi: [10.1155/MPE.2005.101](https://doi.org/10.1155/MPE.2005.101)
- Nadarajah, S. (2005b). Reliability for some bivariate gamma distributions. *Mathematical Problems in Engineering*, 2005(2), 151-163. doi: [10.1155/MPE.2005.151](https://doi.org/10.1155/MPE.2005.151)
- Nadarajah, S., & Kotz, S. (2006). Reliability for some bivariate exponential distributions. *Mathematical Problems in Engineering*, 2006. doi: [10.1155/MPE/2006/41652](https://doi.org/10.1155/MPE/2006/41652)
- Rezaeia, S., Tahmasbib, R., & Mahmoodi, M. (2010). Estimation of $P[Y < X]$ for generalized Pareto distribution. *Journal of Statistical Planning and Inference*, 140(2), 480-494. doi: [10.1016/j.jspi.2009.07.024](https://doi.org/10.1016/j.jspi.2009.07.024)
- Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (2nd ed.). New York, NY: Springer.
- Saraçoğlu, B., & Kaya, M. F. (2007). Maximum likelihood estimation and confidence intervals of system reliability for Gompertz distribution in stress-

strength models. *Selçuk Journal of Applied Mathematics*, 8(2), 25-36. Retrieved from <http://sjam.selcuk.edu.tr/sjam/article/view/188>

Shoukri, M. M., Chaudhary, M. A., & Al-Halees, A. (2005). Estimating $P(Y < X)$ when X and Y are paired exponential variables. *Journal of Statistical Computation and Simulation*, 75(1), 25-38. doi: 10.1080/00949650410001653142

Soliman, A. A., & Al-Aboud, F. M. (2008). Bayesian inference using record values from Rayleigh model with application. *European Journal of Operational Research*, 185(2), 659-672. doi: 10.1016/j.ejor.2007.01.023

Surles, J. G., & Padgett, W. J. (2001). Inference for reliability and stress-strength for a scaled Burr-type X distribution. *Lifetime Data Analysis*, 7(2), 187-200. doi: 10.1023/A:1011352923990

Upadhyay, S. K., & Peshwani, M. (2003). Choice between Weibull and lognormal models: A simulation based Bayesian study. *Communication in Statistics – Theory and Methods*, 32(2), 381-405. doi: 10.1081/STA-120018191

Wang, L., & Shi, Y. (2010). Empirical Bayes inference for the Burr model based on records. *Applied Mathematical Sciences*, 4(34), 1663-1670. Retrieved from <http://www.m-hikari.com/ams/ams-2010/ams-33-36-2010/wangliangAMS33-36-2010.pdf>

Factorial Invariance Testing under Different Levels of Partial Loading Invariance within a Multiple Group Confirmatory Factor Analysis Model

Brian F. French

Washington State University
Pullman, WA

Holmes Finch

Ball State University
Muncie, IN

Scalar invariance in factor models is important for comparing latent means. Little work has focused on invariance testing for other model parameters under various conditions. This simulation study assesses how partial factorial invariance influences invariance testing for model parameters. Type I error inflation and parameter bias were observed.

Keywords: Latent variable modeling, invariance, factor analysis

---

Introduction

Confirmatory factor analysis (CFA) is generally considered the preferred factor analytic approach for assessing scale dimensionality when both theory and empirical evidence support a particular latent structure. CFA is a model-based approach to examining whether there is empirical support for a theoretical latent structure and if the factor structure is equivalent across groups. Questions related to model equivalency across groups falls under the category of measurement invariance (MI; Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Millsap, 2011). A lack of MI is present when an assessment is used to measure a psychological (e.g., motivation) or educational (e.g., mathematical) ability and that assessment produces different results (i.e., scores) for persons from different groups (e.g., boys vs. girls) when those persons are of equal status on the ability assessed (Bollen, 1989; Drasgow & Kanfer, 1985; Millsap, 2011). Stated another way, the measurement properties of the instrument in relation to the ability assessed are the

---

Dr. French is a Professor of Measurement and Psychometrics in the Department of Educational Leadership and Counseling Psychology. Email him at: frenchb@wsu.edu.

Dr. Finch is a Professor of Psychometrics and Statistics in the Department of Educational Psychology. Email him at: whfinch@bsu.edu.

same across pre-identified groups. Methods for identifying a lack of measurement invariance are well-studied. However, the influence of partial invariance, explained below, is not well-documented and becomes difficult to show analytically with many variables (Millsap, 2011). Thus, the goal of the current simulation study is to examine the impact of partial invariance on the invariance testing of factor model parameters (i.e., testing for factorial invariance), including factor intercepts, error variances, factor variances, factor covariances, and factor means under a variety of conditions, including differences in sample size, number of factors, and number of indicators per factor.

Factorial Invariance

Within the measurement invariance area, we focus our emphasis on factorial invariance (FI) examined through the use of multigroup CFA. FI has received increased attention in the past few years with sections of books devoted to the topic (e.g., Millsap, 2011; Schriesheim & Neider, 2001) as well as software being automated (e.g., MPLUS 7.2) to make the assessment of invariance accessible to a wide audience. This emphasis is a direct reflection of the essential role that assessment scores play in society (e.g., high-stakes decisions) ranging from education (e.g., teacher evaluations; international student achievement comparisons) to business (e.g., job applicant decisions). In fact, there is an increasing body of evidence that suggests that many observed differences that are cross-cultural may be contaminated by artifacts of measurement or a lack of factorial invariance (Baumgartner & Steenkamp, 2001; Church et al., 2011; Javaras & Ripley, 2007; Poortinga, 1989). These differences may be related to content meaning issues, translation problems, or even response style differences, and in turn, can result in incorrect decisions regarding individuals from different groups as well as group comparisons (French & Finch, 2008b; Millsap & Kwok, 2004; Steinmetz, 2013).

The examination of the internal structure of the instrument for FI is an important step in providing psychometric evidence supporting the validity argument for score use (American Educational Research Association et al., 2014; Wu, Li, & Zumbo, 2007). FI refers to the situation in which the latent factor structure underlying a scale is equivalent across predefined groups (Meredith & Millsap, 1992; Millsap, 2011; Millsap & Kwok, 2004). FI can be further described in terms of the factor model parameters being equivalent across groups. The investigation of invariance is dependent on a specific variable of interest that separates the groups (e.g., biological sex). This latter definition implies that several

parameters (e.g., factor loadings, intercepts, error variances) are equal across groups. We define the different levels of invariance in the factor model below to make explicit how levels of FI relate to one another.

Given that FI refers to a set of assumptions regarding the invariance of various parameters associated with the factor structure of an instrument, it is important to understand each aspect of FI. Millsap (2011) provides an excellent discussion of the levels of FI from weak, pattern, or metric invariance, which refers to pattern matrix invariance (Horn & McArdle, 1992; Millsap, 2011; Widaman & Reise, 1997) to strong or scalar factorial invariance (SI) referring to factor model intercepts being equal across groups (e.g. Steenkamp & Baumgartner, 1998). The method for assessing the invariance assumptions is based upon multiple groups confirmatory factor analysis (MGCFA) which begins by ensuring the general or configural form (CI) of the factor model is present across groups, where only the number of latent variables present, and the correspondence of observed indicators to factors is the same for all groups in the population. The weak and strong forms of invariance place corresponding constraints on the model. The presence of FI implies that the latent variables are being measured in the same way for the population subgroups under consideration (Wicherts & Dolan, 2010). Said another way, “the question of factorial invariance concerns the extent to which the factor structure underlying the measured variables is the same across multiple populations” (Millsap, 2011, p. 73). This implies that scores on the observed manifestation of the latent variable (i.e. expected score on the scale) are the same for members of different groups who have the same level of the latent trait being measured (Wicherts & Dolan, 2010).

When scalar variance does not hold, it is difficult, if not impossible, to know the extent to which group differences on a mean scale score are due to group differences on the latent trait of interest, or due to group differences on the intercepts (Steinmetz, 2013). The dependency of the differences in scale score means on a lack of invariance exists in applied studies (e.g., French & Mantzicopoulos, 2007). In addition to the three forms of invariance mentioned, it is also possible to assess whether there is group invariance with respect to the unique indicator variance (δ). This strict factorial invariance (SFI; Millsap, 2011) occurs when the factor loadings, intercepts, and unique variances are invariant. In addition, SFI is necessary in order to attribute group differences in the mean and covariance structure of the observed indicators to corresponding differences at the latent variable level (Millsap, 2011).

Multiple Group Confirmatory Factor Analysis (MGCFA)

FI can be assessed using MGCFA, where the standard CFA model is expressed as (Bollen, 1989):

$$x_g = \tau_g + \Lambda_g \xi_g + \delta_g \quad (1)$$

It is possible to have indicators (x_g), intercepts (τ_g), loadings (Λ_g), and unique variances (δ_g) that are specific to each group within the population. Likewise, the indicator covariance matrix and associated latent means can be expressed respectively as:

$$\Sigma_g = \Lambda_g \Psi_g \Lambda_g' + \Theta_g, \quad \mu_g = \tau_g + \Lambda_g K_g \quad (2)$$

such that groups are allowed unique observed covariance matrices (Σ_g), factor loadings (Λ_g), factor covariance matrices (Ψ_g), and unique error matrices (Θ_g). In addition, the observed mean for group g is also a function of the intercept for that group (τ_g), the loadings, and the factor mean (K_g). This implies the factor model holds in each population (Millsap, 2011).

MGCFA can be used to test each level or constraint on the factor model to evaluate FI that was described previously, using a series of nested models. For example, to assess CI, a model is fit such that the number of factors is the same across groups, as are the indicators associated with each of these factors. The model specification allows intercepts, loadings, and unique variances to vary across groups. Good model fit, based on appropriate indices (e.g., chi-square, CFI; Hu & Bentler, 1999), would indicate the presence of CI, and is necessary before the investigation of other types of invariance and placing additional constraints on the model. Often model comparison is made using a difference in chi-square (χ^2) statistic values between the less and more restrictive values. The use of the χ^2 statistic is somewhat problematic as a measure of absolute model fit (Bollen, 1989). However, there is evidence that the $\chi^2_{\text{difference}}$ statistic is an accurate tool for comparing the fit of two nested models (e.g., French & Finch, 2006). Other fit indices have been suggested (e.g., change in CFI, RMSEA) to assess FI. Given the lack of clear guidelines on the accuracy of the amount of change needed to indicate differences, the chi-square statistic remains the focus of this study.

Partial Factorial Invariance

Although FI is desirable for educational and psychological scales to possess, in practice it may be a rare commodity (Millsap & Meredith, 2007). For example, it seems with various types of scales it is unusual for complete FI to exist (Church et al., 2011; French & Gotch, 2013; French & Mantzicopoulos, 2007). Group equality on some but not all factor parameters is known as partial factorial invariance (PI) and does exist on major instruments such as intelligence measures (Maller & French, 2004). In one of the first studies to describe PI, Byrne et al. (1989) explored how researchers identify specific factor parameters (e.g. loadings) that are not group invariant after an initial rejection of the complete invariance hypothesis. Using this sensitivity analysis approach, it is possible to identify and release specific model parameters that differ across groups, leading to a PI model (Millsap, 2011). If PI is indeed found, the next question for researchers is to determine whether these differences in measurement structure are meaningful in practice. While the question of whether or not invariance holds can be addressed in a more or less straightforward manner using the MGCFA methodology described above, the issue of what to do about PI is not so clearly addressed, nor is the impact of PI at one level on assessing invariance at another level well understood.

Goals of the Current Study

The use of MGCFA for testing FI and latent mean differences has experienced growth (Vandenberg & Lance, 2000), with a focus on appropriate practices for testing invariance with attention on accuracy (French & Finch, 2006; Meade & Lautenschlager, 2004; Yoon & Millsap, 2007). Much of this work has stemmed from recommendations for researchers and practitioners not to assume the universal accuracy of MGCFA across many data conditions. While this recommendation has been followed with the implementation of Monte Carlo studies (French & Finch, 2006; Meade & Lautenschlager, 2004; Yoon & Millsap, 2007), gaps remain in the research on several issues related to FI and MGCFA. In particular, MGCFA procedures for identification of a lack of FI require further examination to evaluate accuracy under various conditions where no solution may be fully known with analytic work. There remains uncertainty as to the influence of PI on the assessment of invariance of factor model parameters, including intercepts, error variances, factor variances, and factor means.

There is evidence in the latent variable modeling literature that of the presence of non-invariance for one model parameter can lead to inflated Type I error rates for detection of non-invariance for another model parameter. In the context of IRT,

for example, group differences in item difficulty parameters are associated with inflated Type I error rates for the detection of group differences on the item discrimination parameter (French & Finch, 2008b), even in the presence of no population differences in discrimination parameters across groups. Similarly, when item discrimination values differ across groups, the Type I error rate for detecting group differences in item difficulty values are inflated. These findings have been documented in simulation work, by reviewers of such work, and in applied analysis (Finch & French, 2008a). Given the close link between IRT and CFA models (e.g. McDonald, 1999), the same confound may be present for MGCFA analysis. That is, group difference in factor loadings could lead to inflation of the Type I error rate for testing group differences on factor intercepts, error variances, factor variances, factor covariances, and factor means. Such inflated Type I error rates may in turn be especially problematic for specification searches (Millsap, 2011) to identify true group differences on CFA model parameters. Ideally, an analytic solution to address this issue could be employed. However, as noted by Millsap and Kwok (2004) deriving such an analytic solution becomes exceedingly difficult for models that consist of more than one latent variable and a small number of indicators, leading to the necessity of simulation research.

Given that most real world applications of CFA involve models with multiple factors and multiple indicators, an analytic solution to investigate the impact of non-invariant factor loadings on testing invariance for other model parameters will likely be too limited in scope to be informative for most applications. Thus, we turn to simulations to observe whether the presence of non-invariant loading parameters in a CFA model impacts the testing of invariance for other model parameters as has been reported for a similar situation in the context of IRT (Finch & French, 2008a). Moreover, the examination of bias of parameter estimates in such situations is difficult to derive analytically, whereas through simulation we can determine how PI influences bias (Boomsma, 2013).

Thus, the goal of this study was to begin providing insight to the impact of PI on FI assessment by addressing two research questions:

1. What is the influence of partial factor loading invariance on Type I error and power rates for invariance testing of other model parameters beyond the factor loadings? Specifically, what is the influence of incorrectly modeling such partial invariance? We hypothesize that incorrectly modeling or ignoring factor loading differences across groups will result in inflated Type I error rates when testing the invariance of other model parameters.

2. What influence does partial factor loading invariance have on estimation of other factor model parameters, including intercepts, error variances, factor variances, factor covariances, and factor means? We hypothesize that the estimates of other model parameters will be attenuated for the group with the larger values when groups' factor loading differences are ignored.

Figure 1. Example model used to simulate the data

Methods

To test our hypotheses, a Monte Carlo Simulation study (1000 replications per combination of conditions) was conducted. All simulations were conducted using Mplus 7.1 (Muthén & Muthén, 2013). An example of the model used to simulate the data appears in Figure 1. This example is the simplest model used, with 2 factors and 3 indicators per factor. Two groups were simulated across all conditions. The manipulated variables are described below. These conditions were completely crossed with one another, yielding a total of 240 different simulated conditions, or

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

design cells. In essence, each model parameter for which invariance was assessed can be viewed as representing a unique simulation study. The parameters that were tested for differences in the population were the factor intercepts, error variances, factor variances, factor covariances, and factor means. For each of these parameters group differences were simulated at varying levels, which are described below. In addition, when assessing group invariance for each of these parameters, factor loading values were allowed to vary at different levels. The experimental factors are described below. We also include sample Mplus code for the models in the appendix.

Experimental Factors Manipulated in the Simulation

Percent of factor loading PI To assess the influence of factor loading PI on assessment of other model parameters, loadings were simulated to differ between the groups. Specifically, in one condition loadings were simulated to be equal across groups, while in a second case, they were simulated to differ by 0.25, 0.50, and 1.00. The latter two conditions were included in order to assess the performance of the MGCFA model for invariance testing in more extreme cases of factor noninvariance. For each noninvariant condition, 34% of loadings lacked invariance. This allowed for the conditions where these differences could be ignored to examine what occurs in the case of incorrect modeling of PI or correct modeling of PI. Such conditions can occur with software with automatic testing routines with certain models (e.g., Mplus Analysis = Configural, Metric, Scalar with cross-loadings). Finally, in order to assess the performance of the MGCFA in extreme cases of noninvariance, 68% of the factor loadings were allowed to differ between groups.

Modeling of factor loading noninvariance The modeling of factor loading noninvariance was either correctly specified or incorrectly specified. Correct modeling meant that when the loadings were invariant, they were modeled to be so, whereas when they were not invariant they were correctly modeled to be noninvariant. Finally, incorrect modeling meant that when the factor loadings were simulated to be noninvariant, they were incorrectly modeled as being invariant. Again, this could be a result of Type II errors, automatic software routines, or direction of invariance testing.

Model parameter group differences For each model parameter tested for invariance, several conditions were used for group differences. The

intercept and error variances were simulated to be different across groups on the same indicator variables for which the factor loadings were simulated to be different. All differences were unidirectional (e.g., favoring one group). The intercept values and factor means differed by 0, .2, .5 or .8, representing no difference to a large difference (Cohen, 1988). These were standardized value differences. That is, intercept and latent mean differences are absolute differences whereby one group had a value of 0 for the intercept or factor mean, and the other group had a value of 0.0, 0.2, 0.5, or 0.8. The covariances between factors had standardized values (i.e., correlations) of either 0.0, 0.1, 0.3, or 0.5, representing differences. The factor variance for one group was set to 1 across conditions. The variance of the second group was then varied from 1 (factor variance invariance), 1.33, 1.66, and 2.00, in order to reflect different levels of factor variance noninvariance. In the simulated models, the assumption was made that there was no specific variance (i.e. all error variance) and the value of the theta-deltas was 1.0 minus the square of the respective factor loading, as we set the loading values. This minimized potential confounding factors in examining the results. The differences simulated reflect a range of values that represent typical small to large differences. These were not tied to any content area as Cohen (1988) suggests but were broad strokes to capture situations that could be applied to many areas of work. In other words, the goal was to study a range of potential group parameter differences from none through moderate and large.

Sample size The total sample was simulated to be 300, 1000, and 2000 with equal group sizes. These values are designed to represent cases from the smallest samples generally seen in practice (French & Mantzicopoulos, 2007), to what would be considered a large sample in most social science applications and common conditions in simulation work (French & Finch, 2006; 2008b; Meade & Lautenschlager, 2004; Steinmetz, 2013) while maintaining adequate statistical power (Hancock & French, 2013). The smallest total sample size used here was 300, meaning that each group contained 150 individuals. Samples smaller than that were not used as it could lead to unstable parameter estimates, particularly for the more complex models (Kline, 2011), and confound the results.

Number of factors and indicators per factor The number of indicators per factor simulated was 3 or 6 which were completely crossed with the number of factors of either 2 or 4. This range of values is designed to reflect both extensively measured and less extensively measured constructs, and was in accord with prior research in this area (e.g. Millsap & Kwok, 2004; Steinmetz, 2013) to

facilitate generalizability. In addition, the models used in this simulation study are similar to models published in actual practice (e.g. Bavarian et al., 2014; Hesse & Klingberg, 2014; Tam, 2014).

Percent of invariance in factor intercepts, error variances, factor variances, factor covariances, and factor means

Three levels of the amount of non-invariance across groups were simulated. To assess Type I error (i.e., false identification of a lack of invariance) of the MGCFA methods employed, the case of complete invariance (i.e. no differences in model parameters across groups) was simulated. In addition, to assess power (i.e., correct identification of a lack of invariance) 34% of target model parameters lacked invariance.

Response variables

Several outcome variables were examined including Type I error rates and power for the chi-square difference test χ^2_{Δ} at $\alpha = 0.05$, parameter estimation bias (sample parameter estimate – population parameter value), standard deviation of the parameter estimates, mean square error (MSE) for parameter estimates, and parameter coverage rates for 95% confidence intervals. To determine which of the manipulated variables or their interactions significantly impacted the Type I error and power rates, analysis of variance (ANOVA) was used (Paxton, Curran, Bollen, Kirby, & Chen, 2001), in addition to the effect size measure (η^2) to assist with identifying effects worth noting. The outcome variable for the ANOVA model was the total number of the 1000 replications for each combination of conditions in which the null hypothesis of invariance was rejected. With regard to criteria for acceptable performance, Type I error rates were considered acceptable if they were at the nominal 0.05 level, just as coverage rates for parameter estimates were acceptable when the actual coverage was at the nominal 0.95 value. For power, we considered the typical value of 0.80 to be the criterion.

Results

Model Parameter Invariance: Type I Error and Power when Factor Loadings Differed by 0.5 or 1.0

Following are the results for Type I error and power rates for assessing the invariance assumption for the various model parameters. As noted above, simulations were conducted for factor loading differences between groups of 0.25, 0.50, and 1.00. Results showed that the Type I error rates for assessing invariance

of other model parameters when factor loadings differed between the groups by 0.50 and 1.00 were highly inflated when the lack of loading invariance was not properly modeled. When group loading differences were not properly modeled and the loadings differed by 0.50 or 1.00, the Type I error rates for incorrectly identifying noninvariance for factor intercepts, error variances, factor variances, factor covariances, and factor means were at or above 0.80 in the 34% noninvariant case, and at or above 0.95 in the 68% noninvariant case. These inflated rates were present regardless of the other manipulated conditions, including sample size, number of indicators, and number of factors. On the other hand, when the lack of factor loading invariance was correctly modeled, the Type I error rates for assessing invariance of the other model parameters were between 0.045 and 0.058 for all model parameters, across sample size, number of factors, and number of indicators. Given the uniformly inflated Type I error rates for the incorrectly modeled factor loading noninvariance condition, power for this case was not investigated for any of the model parameters, as these rates cannot be interpreted with any confidence.

When the factor loadings were correctly modeled, power rates in the 0.50 and 1.00 factor loading difference cases were very similar to power rates in the 0.25 factor loading difference condition. Therefore, in order to save space, we report only the power values for the 0.25 factor loading difference condition in the following section of the paper. In addition, given that the Type I error rates were extremely inflated when the factor loadings were simulated to differ between groups by 0.50 and 1.00, and that they were essentially identical to those obtained when the loadings differed by 0.25 and this lack of invariance was correctly modeled, it was felt that reporting results for the two larger noninvariant conditions would be redundant. Therefore, the results that appear below reflect only the cases where the loadings were truly invariant, or where they differed between the groups by 0.25.

Intercept Invariance: Type I Error and Power

The ANOVA for the Type I error rate reveal no significant interactions or main effects ($p = 0.05$). The Type I error rate, parameter bias for the indicators on which intercepts were simulated to differ, parameter standard deviation (SD), mean squared error (MSE), and coverage rates appear in [Table 1](#). Across conditions, the Type I error rate for χ^2_A was at the nominal 0.05 rate. Parameter estimates were somewhat negatively biased, and coverage rates were close to the nominal 0.95 both when the intercepts were constrained to be equal and when they were not. In addition, the SD and MSE of the parameter estimates were very comparable under

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

both unconstrained and constrained conditions. Notably given the goals of this study, the Type I error rate was not influenced by factor loading PI, nor by whether that PI was correctly or incorrectly modeled.

The ANOVA identified sample size (N) as being significantly related to power rates ($F_{2,88} = 8.4$, $p < 0.001$, $\eta^2 = 0.16$), as well as degree of intercept difference ($F_{2,88} = 41.8$, $p < 0.001$, $\eta^2 = 0.59$). Power rates increased from 0.69 for a total sample size of 300, to nearly 1.00 for samples of 1000 and 2000. Table 1 includes the power, bias, SD, MSE, and coverage rates for the unconstrained and constrained modeling conditions, by the degree of intercept difference.

Table 1. Intercept invariance testing Type I error rate and power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept across simulated conditions: Unconstrained parameters/Constrained parameters

Difference	Type I error	Bias	SD	MSE	Coverage
0.0	0.051	-0.01 / -0.01	0.06 / 0.05	0.005 / 0.004	0.93 / 0.93
	Power				
0.2	0.780	0.03 / -0.08	0.06 / 0.05	0.010 / 0.020	0.93 / 0.92
0.5	0.990	0.05 / -0.15	0.06 / 0.05	0.030 / 0.050	0.92 / 0.66
0.8	1.000	0.05 / -0.25	0.06 / 0.06	0.030 / 0.110	0.93 / 0.54

Power for detecting intercept differences increased concomitantly with increases in the population difference between the groups' intercept values, which would be expected. In addition, for the group with the larger intercept when the intercept estimates were simulated to differ, the parameter estimate displayed greater bias than when no group differences were simulated (Table 1). For the unconstrained condition, there was a positive bias in the intercept estimate for the group with the larger intercept, while for the constrained condition there was negative bias that increased with greater group differences in the population intercept value. This result was expected for the constrained condition because the groups' intercepts were forced to be equal, thus driving down the value for the group with the larger intercept. In addition, whereas the SD of the estimates was comparable for both conditions across the size of intercept difference, the MSE increased and coverage decreased with greater such differences in the constrained condition but remained largely unchanged in the unconstrained case. Additionally, as was true with the Type I error, the loading PI condition, along with how it was modeled, had no impact on the assessment of intercept differences.

Table 2. Error variance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for error variance by partial loading invariance and modeling conditions, and by number of factors and indicators per factor: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	0.002 / 0.002	0.01 / 0.01	0.001 / 0.001	0.94 / 0.94
Partial correct	0.05	-0.002 / 0.002	0.01 / 0.01	0.001 / 0.001	0.96 / 0.94
Partial incorrect	0.44	-0.010 / -0.020	0.01 / 0.01	0.010 / 0.010	0.62 / 0.57

Correct modeling of partial loading invariance or Full loading invariance					
Factors / indicators per factor	Type I error	Bias	SD	MSE	Coverage
2 / 3	0.05	0.002 / 0.002	0.01 / 0.01	0.0020 / 0.0010	0.94 / 0.95
2 / 6	0.05	0.002 / 0.002	0.01 / 0.01	0.0010 / 0.0010	0.95 / 0.95
4 / 3	0.05	-0.010 / 0.002	0.01 / 0.01	0.0050 / 0.0010	0.78 / 0.94
4 / 6	0.05	0.002 / 0.002	0.01 / 0.01	0.0003 / 0.0003	0.94 / 0.94

Incorrect modeling of partial loading invariance					
Factors / indicators per factor	Type I error	Bias	SD	MSE	Coverage
2 / 3	0.34	0.005 / -0.001	0.01 / 0.01	0.003 / 0.002	0.93 / 0.94
2 / 6	0.18	0.000 / -0.002	0.01 / 0.01	0.001 / 0.001	0.95 / 0.92
4 / 3	0.98	-0.040 / -0.060	0.01 / 0.01	0.020 / 0.040	0.08 / 0.02
4 / 6	0.27	-0.010 / -0.010	0.01 / 0.01	0.002 / 0.002	0.53 / 0.41

Error Variances: Type I Error and Power

The ANOVA identified the interaction between factor loading difference and modeling of that difference ($F_{1,88} = 16.8, p < 0.001, \eta^2 = 0.51$), and the interaction of the number of factors by number of indicators per factor ($F_{1,88} = 6.2, p = 0.024, \eta^2 = 0.28$) as significantly related to the Type I error rate for detecting differences in group error variances. Table 2 contains Type I error rates, parameter bias, SD, MSE, and coverage rates by loading PI and modeling conditions. These results show that when the loadings are fully invariant, or PI with the invariance being correctly modeled, the Type I error rates are at the nominal 0.05 level. However, when the loadings are PI but modeled as fully invariant, the Type I error rate for testing error variance was inflated to 0.44. A further examination of the results in Table 2 reveals that under the fully invariant or partial correct conditions, parameter bias, SD, MSE are all relatively low, and the coverage rates are at the nominal level for both the unconstrained and constrained models. However, parameter bias was 5 times larger for the unconstrained model and 10 times larger for the constrained model in the partial incorrect condition, while the coverage rates for both modeling conditions was well below the nominal 0.95 level.

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

In addition, these results in Table 2 are divided into those in which full factor loading held, or loading PI was correctly modeled, and those in which loading PI was not correctly modeled. The Type I error rate was found to be at the nominal (0.05) level when loadings were fully invariant or PI was correctly modeled. However, when loadings were PI but not correctly modeled the error rates were inflated, with greater inflation occurring for 3 indicators per factor. In addition, the greatest inflation occurred when there were 4 factors each with 3 indicators. A condition for which parameter estimate bias was also greatest, and coverage was lowest.

Error Variances: Power

The ANOVA indicated that none of the manipulated factors were significantly related to the power to detect error noninvariance. Across conditions, power for detecting error noninvariance was extremely high (0.99). There was a much larger negative bias for the constrained parameter model than in the unconstrained case (-0.060 vs -0.003). In addition, the MSE was more than 10 times larger for the constrained model, and displayed coverage rates of just 0.10, well below the nominal 0.95 level. For the unconstrained model, the parameter coverage rate was also below the nominal level, at 0.83, but much higher than for the constrained model. Of particular interest in this study, power rates were not significantly influenced by PI of the factor loadings, unlike in the Type I error case. This result would appear to be in large part due to the extremely high power for detecting error noninvariance across conditions.

Factor Covariance: Type I Error and Power

With respect to the Type I error rate when testing the invariance hypothesis of factor covariances, the ANOVA results showed that the interaction of factor loading difference and the modeling of that difference ($F_{2,35} = 8.8$, $p = 0.001$, $\eta^2 = 0.35$) was statistically significant. Table 3 includes the outcome variables of interest by the factor loading difference and modeling of the difference, and by the number of factors. The Type I error rate showed some inflation when there was loading PI that was not properly modeled. Accompanying this Type I error inflation was greater negative bias of parameter estimates, particularly for the constrained parameter model, which in turn was associated with inflation of the MSE, again particularly in the case of incorrect modeling of the factor loading PI. Finally, coverage rates were lower in the PI incorrect condition for both the constrained and unconstrained model, with coverage higher for the unconstrained model.

The ANOVA for power in testing covariance parameter invariance identified sample size as the only statistically significant variable ($F_{2,35} = 5.7$, $p = 0.004$, $\eta^2 = 0.12$). Power rates, bias, SD, MSE, and coverage rates appear in Table 4. Power increased from 0.61 for $N = 300$ to 0.88 for $N = 2000$. Parameter bias, MSE, and coverage rates were largely unaffected by sample size, though SD for both the constrained and unconstrained models was lower for the two larger sample sizes than for $N = 300$. In addition, bias was lower for the unconstrained model as compared to the constrained model, and MSE and coverage rates were higher. Power for detecting noninvariant factor covariances was not found to be influenced by loading PI or how it was modeled.

ANOVA did not identify any significant effects for the Type I error rate in the detection of factor mean differences between groups. In this case, bias, SD, MSE, and coverage rates appear only for the unconstrained model because in the constrained case, both groups' means were set equal to 0. The Type I error rate for testing group mean invariance was at the nominal 0.05 rate, with low bias for the mean that was allowed to vary, and a coverage rate at the nominal 0.95 level. Of particular interest was the fact that the loading PI condition and the way that this was modeled, were not significantly related to the Type I error rate when testing for factor mean differences between groups.

Table 3. Factor covariance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	-0.0001 / -0.0900	0.07 / 0.05	0.12 / 0.19	0.96 / 0.82
Partial correct	0.05	0.0002 / -0.1000	0.07 / 0.05	0.19 / 0.25	0.93 / 0.80
Partial incorrect	0.07	-0.0003 / -0.2100	0.07 / 0.07	0.19 / 0.37	0.91 / 0.71

Table 4. Factor covariance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by sample size (N): Unconstrained parameters/Constrained parameters

N	Power	Bias	SD	MSE	Coverage
150 / 150	0.61	-0.002 / -0.130	0.10 / 0.09	0.20 / 0.24	0.94 / 0.83
500 / 500	0.76	-0.004 / -0.140	0.06 / 0.05	0.19 / 0.24	0.94 / 0.80
1000 / 1000	0.88	0.007 / -0.140	0.04 / 0.04	0.19 / 0.24	0.95 / 0.82

Factor means: Type I error and Power

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

Table 5. Factor mean invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by sample size (N): Unconstrained parameters

N	Power	Bias	SD	MSE	Coverage
150 / 150	0.78	-0.015	0.14	0.35	0.93
500 / 500	0.95	-0.013	0.07	0.33	0.94
1000 / 1000	0.99	-0.013	0.05	0.30	0.93

Difference	Power	Bias	SD	MSE	Coverage
0.2	0.73	-0.006	0.09	0.05	0.96
0.5	0.99	-0.014	0.09	0.27	0.94
0.8	1.00	-0.038	0.09	0.71	0.90

The ANOVA results showed that the sample size ($F_{2,91} = 22.3$, $p < 0.001$, $\eta^2 = 0.33$) and the degree of factor mean difference ($F_{2,91} = 46.4$, $p < 0.001$, $\eta^2 = 0.50$) were significantly related to power rates. Table 5 includes power, bias, SD, MSE, and coverage rates by sample size, and group factor mean difference, for the unconstrained model only. Power increased concomitantly with sample size, as would be expected. Furthermore, power was above 0.75 in the worst case, and at 0.95 or above for samples of 1000 or more. Parameter bias appears to not have been influenced by sample size, though the SD and MSE declined somewhat with increasing sample sizes. Coverage rates were near the nominal 0.95 level across sample sizes.

With respect to the difference between group factor means, power rates exceeded 0.7 even for the smallest difference of 0.2. There was an increase in the amount of negative bias as the group mean difference increased, indicating that, for the group whose mean was larger, the unconstrained model provided a slight underestimate. It should be noted, however, that the largest bias value was -0.038, indicating that the factor mean estimate was approximately 0.76 when in the population the value was 0.80. This increase in bias with a greater group mean difference was also associated with an increase of MSE and a decrease in the coverage rate to 0.90, below the nominal 0.95 rate. The power for testing group factor mean differences was not significantly influenced by loading PI or whether it was modeled correctly.

Factor Variances: Type I Error and Power

ANOVA results showed that the interaction between loading PI and its modeling was the only term significantly related to the Type I error rate for testing factor variance invariance ($F_{2,36} = 22.7$, $p < 0.001$, $\eta^2 = 0.59$). Table 6 includes the Type

I error rates, bias, SD, MSE, and coverage rates for the unconstrained and constrained models when testing factor variance invariance. The Type I error rate for testing group variance invariance was at the nominal rate when full factor loading invariance held, or when there was loading PI and it was modeled correctly. However, when the loadings were PI but modeled as fully invariant, the Type I error rate was inflated to 0.30. In addition, bias in the factor variances for both the unconstrained and constrained models was much lower in the fully invariant or partial correct conditions, than in the partial incorrect. This increase in bias was associated with inflated MSE and lower coverage rates for both the constrained and unconstrained models, though SD was not impacted by the loading invariance condition.

The ANOVA results showed that the 3-way interaction of N by loading PI by modeling of PI ($F_{2,86} = 53.4$, $p < 0.001$, $\eta^2 = 0.55$), and the main effect of factor variance difference ($F_{4,86} = 4.7$, $p = 0.012$, $\eta^2 = 0.10$) were the only statistically significantly related terms to the power for detecting noninvariant group variances. Power, bias, SD, MSE, and coverage rates by N , loading invariance, and modeling conditions appear in Table 7. Power increased with sample size and was slightly higher (approximately 0.04) in the full versus partial invariance conditions for the two smaller sample sizes. Power for the PI condition modeled incorrectly should not be interpreted given the observed inflated Type I error rate. With regard to the estimated factor variances, larger positive bias was observed in the unconstrained group in the incorrect PI condition than either when full invariance held, or in the correctly modeled PI condition. This positive bias indicates that the variance for the group simulated to have the larger value was overestimated when the factor loadings were noninvariant but constrained to be equal across groups.

On the other hand, for the constrained model (where variances for the two groups were held equal), the variance estimate for the group with the larger population value was negatively biased. The MSE for the constrained group was

Table 6. Factor variance invariance testing Type I error rate, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

Loading invariance	Type I error	Bias	SD	MSE	Coverage
Full	0.05	-0.002 / 0.004	0.14 / 0.14	0.15 / 0.16	0.94 / 0.92
Partial correct	0.05	-0.001 / 0.002	0.15 / 0.14	0.15 / 0.16	0.94 / 0.93
Partial incorrect	0.30	0.080 / -0.069	0.14 / 0.13	0.21 / 0.20	0.77 / 0.72

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

Table 7. Factor variance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for testing covariance invariance by sample size (N), partial loading invariance and modeling conditions: Unconstrained parameters/Constrained parameters

N	Loading Invariance	Power	Bias	SD	MSE	Coverage
150 / 150	Full	0.63	0.001 / -0.308	0.21 / 0.21	0.25 / 0.46	0.95 / 0.53
	Partial correct	0.59	-0.033 / -0.337	0.22 / 0.21	0.26 / 0.46	0.93 / 0.50
	Partial incorrect	0.83*	0.103 / -0.359	0.20 / 0.21	0.36 / 0.47	0.86 / 0.55
500 / 500	Full	0.91	0.003 / -0.328	0.11 / 0.12	0.25 / 0.47	0.94 / 0.51
	Partial correct	0.87	-0.002 / -0.340	0.11 / 0.12	0.24 / 0.46	0.94 / 0.52
	Partial incorrect	0.98*	0.097 / -0.361	0.12 / 0.11	0.36 / 0.47	0.87 / 0.54
1000 / 1000	Full	0.97	0.003 / -0.325	0.08 / 0.08	0.24 / 0.46	0.93 / 0.55
	Partial correct	0.96	0.003 / -0.338	0.08 / 0.08	0.24 / 0.47	0.95 / 0.54
	Partial incorrect	0.99*	0.105 / -0.362	0.07 / 0.08	0.35 / 0.48	0.86 / 0.54

* Power rates in bold are associated with conditions in which the Type I error rate was inflated

Table 8. Factor variance invariance testing power, parameter bias, parameter estimate standard deviation, MSE, and coverage rates for intercept by intercept difference: Unconstrained parameters/Constrained parameters

Difference	Power	Bias	SD	MSE	Coverage
0.33	0.70	0.03 / -0.17	0.14 / 0.13	0.28 / 0.39	0.92 / 0.49
0.66	0.91	0.03 / -0.34	0.13 / 0.14	0.28 / 0.45	0.92 / 0.53
1.00	0.98	0.04 / -0.51	0.13 / 0.14	0.29 / 0.53	0.91 / 0.59

greater in the incorrect PI condition than for either full loading invariance or correct modeling of loading PI. However, for the constrained model, there was no notable difference in MSE across these conditions. Parameter coverage was also well below the nominal 0.95 rate for the constrained model, while in the unconstrained case coverage was at or near the nominal rate except when factor loadings were PI but not modeled as such.

Table 8 includes the power, bias, SD, MSE, and coverage rates by level of variance difference. Power for detecting unequal group variances increased as the degree of that difference increased, which would be expected. For the unconstrained model, the amount of bias, the SD, MSE, and coverage rates were essentially the same across differences in group variances. In contrast, for the constrained model, the amount of bias increased concomitantly with differences in the magnitude of the group factor variances. Again, this result is expected when one considers that for the constrained model, one variance is simulated to be 1.00,

while the other is simulated to be 1.33, 1.66, or 2.00 in the simplest two factor case. Thus, when the variances are constrained to be equal, the magnitude of bias should increase along with increases in the difference between group variances.

Discussion

The use of MGCFA for testing FI and latent mean differences will continue to grow as modeling of outcomes over time and across diverse groups takes greater advantage of the rapid methodological changes in latent variable modeling. This increased use leads to a need for focusing on appropriate practices when assessing invariance, especially in the presence of PI, be it for the traditional MGCFA or other models with grouping variables and measurement models (e.g., latent profile analysis). Accurate invariance testing rests upon the assumption that MGCFA works well across many data conditions. Yet, heretofore, empirical research in this area has focused primarily on the performance of tests for factor loading invariance. In particular, there is a pressing need to provide researchers with useful information on how PI affects observed composite scores on assessments. Perhaps more importantly, researchers and practitioners need to be provided assistance in understanding how decisions made about groups and individuals using such scales are impacted by PI. Our current work attempts to address these issues in the FI and MGCFA research domains by examining the accuracy for assessing invariance at all levels of the CFA model across groups under various levels of factor loading invariance. The results allowed us to draw a few main conclusions.

First, when factor loading PI is correctly modeled, invariance testing on the other model parameters was not adversely influenced regardless of how large the group differences in factor loadings were. However, and second, if modeling of such differences is done incorrectly and the degree of group loading difference is 0.50 or 1.00, then invariance testing for other model parameters will suffer from Type I error inflation of 0.8 or higher. When the degree of loading difference is 0.25, and this lack of invariance is not modeled correctly, there will also likely be Type I error inflation for testing the invariance of error variances and factor variances and, to a much smaller extent, factor covariances. Thus, careful attention must be paid to the correct modeling of partial factor loading invariance when researchers are interested in assessing invariance of the latent model variance and covariance structures. Third, PI of the factor loadings with group differences of 0.25 had no impact on testing the invariance of intercepts or factor means, whether the factor loading differences were correctly modeled or not. This result was surprising given the distortion of Type I error in other measurement invariance

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

studies where one level of non-invariance influenced another (French & Finch, 2008b) and the distortion of mean differences under a lack of factorial invariance (Millsap, 2011; Steinmetz, 2013). However, considered in light of the Type I error inflation for mean and intercept differences associated with factor loading differences of 0.50 and 1.00, it would appear that the impact of PI is simply not felt until group differences on the loadings reaches a critical juncture. Fourth, model complexity, defined as the number of factors and indicators, only influenced invariance testing for error variances. Additionally, this was only a concern when the PI of factor loadings was modeled incorrectly. Under such conditions, more factors were associated with greater Type I error rates and bias, and lower coverage.

Taken together, these results demonstrate the importance of assessing and correctly modeling the invariance of factor loadings prior to testing for invariance in other model parameters. However, if this is done and if PI in loadings is modeled correctly, the researcher can be confident that it will not impact assessment of other model parameters, even when one group has a majority of loadings that are twice the size of the other groups' loadings. This is likely the best outcome that can be achieved, and yet one that is the most challenging to achieve as it is difficult to be certain that all such group differences in loadings have been correctly modeled. Thus, the applied researcher must be keenly aware of both the steps that they take in testing parameter invariance and how they account for loading PI. This requires an awareness of the correct sequence of steps used in invariance testing, and a knowledge of what software programs with automated functions are doing to account for different levels of invariance as the program systematically tests for full invariance. This latter issue becomes ever more challenging as new versions of software are released on a yearly basis with increasingly automated functions for conducting invariance testing (e.g., Mplus, IRTPRO). This is not to say such automation is entirely negative. Rather, it is a call for clearer documentation of the steps that are being taken and the underlying assumptions and model constraints that are imposed and, perhaps most importantly, users taking the responsibility to understand what they are modeling.

The purpose of this investigation was to provide evidence of MGCFA performance for FI testing under a variety of practical and applied conditions, specifically focused on models where loading PI is present. However, although we estimated many models under various conditions, simulation of exhaustive conditions is not practically possible. Therefore, additional simulation work is encouraged to continue examining MGCFA analyses under an even greater array of conditions (e.g., percent of misspecification, mixed invariant conditions) as there are several problems which remain to be solved in invariance testing (Millsap,

2011). We note that analytic solutions should be sought first or in conjunction with simulation work to aid the understanding of the underlying models and reasons for differences in outcomes (Boomsma, 2013). That said, hopefully this research will inform practice for those engaged in FI analyses to better understand phenomena in their disciplines. The results described here should allow practitioners to make informed decisions in the presence of loading PI. Furthermore, they should inform practice by highlighting the strengths and limitations of MGCFA given certain conditions. We also think these results can stimulate new research surrounding the implementation of the MGCFA and other latent group models. For example, the development of an effect size to capture the magnitude of the difference in parameters and the influence it has on latent mean difference testing would be helpful in allowing power results to be more meaningfully examined (Millsap, 2011). This line of work should lead to accurate decisions about individuals and groups in the presence of partial invariance through MGCFA analysis.

References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baumgartner, H., & Steenkamp, J. E. M. (2001). Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research*, 38(2), 143-156. doi: [10.1509/jmkr.38.2.143.18840](https://doi.org/10.1509/jmkr.38.2.143.18840)

Bavarian, N., Flay, B.R., Ketcham, P.L., Smit, E., Kodama, C., Martin, M., & Saltz, R.F. (2014). Using structural equation modeling to understand prescription stimulant misuse: A test of the theory of triadic influence. *Drug and Alcohol Dependence*, 138(1), 193-201. doi: [10.1016/j.drugalcdep.2014.02.700](https://doi.org/10.1016/j.drugalcdep.2014.02.700)

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Boomsma, A. (2013). Reporting Monte Carlo studies in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 20(3), 518-540. doi: [10.1080/10705511.2013.797839](https://doi.org/10.1080/10705511.2013.797839)

- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *Psychological Bulletin*, 105(3), 456-466. doi: [10.1037/0033-2909.105.3.456](https://doi.org/10.1037/0033-2909.105.3.456)
- Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology*, 101(5), 1068-1089. doi: [10.1037/a0025290](https://doi.org/10.1037/a0025290)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hoboken, NJ: Lawrence Erlbaum Associates.
- Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70(4), 662-680. doi: [10.1037/0021-9010.70.4.662](https://doi.org/10.1037/0021-9010.70.4.662)
- French, B. F., & Finch, W. H. (2006). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 13(3), 378-402. doi: [10.1207/s15328007sem1303_3](https://doi.org/10.1207/s15328007sem1303_3)
- Finch, W. H., & French, B. F. (2008a). Anomalous Type I error rates for identifying one type of differential item functioning in the presence of another. *Educational and Psychological Measurement*, 68(5), 742-759. doi: [10.1177/0013164407313370](https://doi.org/10.1177/0013164407313370)
- French, B. F., & Finch, W. H. (2008b, March). *When under the influence of non-invariant factor loadings, does computation method of the factor score matter?* Paper presented at the annual meeting of the American Educational Research Association. New York, NY.
- French, B. F., & Gotch, C. M. (2013). Sex differences in item functioning in the comprehensive inventory of basic skills-II vocabulary assessments. *Journal of Psychoeducational Assessment*, 31(4), 410-417. doi: [10.1177/0734282912460857](https://doi.org/10.1177/0734282912460857)
- French, B. F., & Mantzicopoulos, P. (2007) An examination of the first/second-grade form of the pictorial scale of perceived competence and social acceptance: Factor structure and stability by grade and gender across groups of economically disadvantaged children. *Journal of School Psychology*, 45(3), 311-331. doi: [10.1016/j.jsp.2006.07.004](https://doi.org/10.1016/j.jsp.2006.07.004)
- Hancock, G. R. & French, B. F. (2013). Power analysis in structural equation modeling. In G. R. Hancock & R. O. Muller (Eds.), *Structural equation*

modeling: A second course (2nd ed.) (pp. 117-159). Charlotte, NC: Information Age Publishing.

Hesse, K., & Klingberg, S. (2014). Examining the cognitive model of caregiving – A structural equation modeling approach. *Psychiatry Research*, 217(3), 171-176. doi: [10.1016/j.psychres.2014.02.023](https://doi.org/10.1016/j.psychres.2014.02.023)

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, 18(3), 117-144. doi: [10.1080/03610739208253916](https://doi.org/10.1080/03610739208253916)

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1-55. doi: [10.1080/10705519909540118](https://doi.org/10.1080/10705519909540118)

Javaras, K. N., & Ripley, B. D. (2007). An “unfolding” latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association*, 102(478), 454-463. doi: [10.1198/016214506000000960](https://doi.org/10.1198/016214506000000960)

Maller, S. J., & French, B. F. (2004). Universal Nonverbal Intelligence Test factor invariance across deaf and standardization samples. *Educational and Psychological Measurement*, 64(4), 647-660. doi: [10.1177/0013164404263877](https://doi.org/10.1177/0013164404263877)

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 11(1), 60-72. doi: [10.1207/S15328007SEM1101_5](https://doi.org/10.1207/S15328007SEM1101_5)

Meredith, W., & Millsap, R. E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, 57(2), 289-311. doi: [10.1007/BF02294510](https://doi.org/10.1007/BF02294510)

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.

Millsap, R. E., & Kwok, O. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Measurement*, 9(1), 93-115. doi: [10.1037/1082-989X.9.1.93](https://doi.org/10.1037/1082-989X.9.1.93)

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: historical perspectives and new problems. In R. Cudeck & R. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131-152). Mahwah, NJ: Erlbaum.

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

- Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide, version 7.1*. Los Angeles, CA: Muthén and Muthén.
- Poortinga, Y. P. (1989). Equivalence of cross-cultural data: An overview of basic issues. *International Journal of Psychology*, 24(6), 737-756. doi: [10.1080/00207598908247842](https://doi.org/10.1080/00207598908247842)
- Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2), 287-312. doi: [10.1207/S15328007SEM0802_7](https://doi.org/10.1207/S15328007SEM0802_7)
- Schriesheim, C. A., & Neider, L. L. (2001). *Equivalence in measurement*. Greenwich, CN: Information Age.
- Steenkamp, J. E. M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, 25(1), 78-90. doi: [10.1086/209528](https://doi.org/10.1086/209528)
- Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, 9, 1-12. doi: [10.1027/1614-2241/a000049](https://doi.org/10.1027/1614-2241/a000049)
- Tam, F. W. (2014). A structural equation model of drug abuse among secondary students in Hong Kong. *International Journal of Mental Health and Addiction*, 12(3), 255-269. doi: [10.1007/s11469-013-9450-5](https://doi.org/10.1007/s11469-013-9450-5)
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, 3(1), 4-70. doi: [10.1177/109442810031002](https://doi.org/10.1177/109442810031002)
- Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practices*, 29(3), 39-47. doi: [10.1111/j.1745-3992.2010.00182.x](https://doi.org/10.1111/j.1745-3992.2010.00182.x)
- Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle, & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychiatric Association
- Wu, A. D., Li, Z., & Zumbo, B. D. (2007). Decoding the meaning of factorial invariance and updating the practice of multi-group confirmatory factor analysis: A demonstration with the TIMSS data. *Practical Assessment, Research,*

& Evaluation, 12(3), 1-26. Retrieved from
<http://pareonline.net/getvn.asp?v=12&n=3>

Yoon, M. & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A Monte Carlo study. *Structural Equation Modeling*, 14(3), 435-463. doi: 10.1080/10705510701301677

Appendix

Sample Mplus Code

Model lack of invariance

```

title:      Model lack of invariance
data:      file is replist.dat;
           type=montecarlo;
variable:  names are y1-y6 group;
           grouping is group (1=g1 2=g2);

model:
    f1 by y1;
    f1 by y2;
    f1 by y3;
    f2 by y4-y6;
    f1@1 f2@1;
    f1 with f2;
    y1-y6;

model g1:  f1 by y2* ;
           [y1*0] (1);
           [y2*0] (2);
           [y3*] ;
           [y4*0] (4);
           [y5*0] (5);
           [y6*0] (6);

model g2:
    f1 by y2* ;
    [y1*0] (1);
    [y2*0] (2);
    [y3*] ;
    [y4*0] (4);
    [y5*0] (5);
    [y6*0] (6);
savedata:  results are diffresults.out;

```

Model total invariance

```

title:      Model total invariance
data:      file is replist.dat;
           type=montecarlo;
variable:  names are y1-y6 group;
           grouping is group (1=g1 2=g2);
model:

```

FRENCH & FINCH

```
      f1 by y1;
      f1 by y2;
      f1 by y3;
      f2 by y4-y6;
      f1@1 f2@1;
      f1 with f2;
      y1-y6;

model g1:  f1 by y2* ;
           [y1*0] (1);
           [y2*0] (2);
           [y3*0] (3);
           [y4*0] (4);
           [y5*0] (5);
           [y6*0] (6);

model g2:
           f1 by y2* ;
           [y1*0] (1);
           [y2*0] (2);
           [y3*0] (3);
           [y4*0] (4);
           [y5*0] (5);
           [y6*0] (6);

savedata:  results are nodiffresults.out;
```

Test run

```
title:      test run
montecarlo: names are y1-y6;
            nobervations=1000 1000;
            nreps=1000;
            seed=94756;
            ngroups=2;
            repsave=all;
            save=rep*.dat;

model population:
           [y1-y6@0];
           y1-y6@1;
           f1 by y1@1 y2-y3*.6;
           f2 by y4@1 y5-y6*.6;
           f1@1 f2@1;
           f1 with f2@.5;
model population-g2:
f1 by y2*.85;
[y3*0.2];
```

PARTIAL INVARIANCE INFLUENCE ON FACTOR INVARIANCE

```
model:
    f1 by y1;
    f1 by y2;
    f1 by y3;
    f2 by y4;
    f2 by y5;
    f2 by y6;
    f1@1 f2@1 (1);
    f1 with f2 (2);
    y1-y6 (3);

model g2: [y3*0.2];
f1 by y2*.85;
output:    tech9;
```

A Comparison of Estimation Methods For Nonlinear Mixed Effects Models Under Model Misspecification and Data Sparseness: A Simulation Study

Jeffrey R. Harring
University of Maryland
College Park, MD

Junhui Liu
Educational Testing Service
Princeton, NJ

A Monte Carlo simulation is employed to investigate the performance of five estimation methods of nonlinear mixed effects models in terms of parameter recovery and efficiency of both regression coefficients and variance/covariance parameters under varying levels of data sparseness and model misspecification.

Keywords: Random coefficient models, linearization, quadrature, Bayesian, nonlinear models, non-normality

---

Introduction

A common challenge for substantive researchers across numerous research domains is to make inferences on features underlying profiles of continuous repeated measures data for a sample of individuals from a population of interest. Nonlinear mixed effects (NLME) models (Davidian & Giltinian, 1995; Pinheiro & Bates, 2000; Vonesh & Chinchilli, 1997) have become the tools of choice for analyses in which the primary interest of researchers focuses on understanding the nature of systematic and random variation between and within individuals. The biomedical literature, for example, is replete with studies from areas like pharmacokinetics, which have developed NLME models to examine drug concentration and dispersion in patients (see e.g., Beal & Sheiner, 1985) or modeling markers of disease progression (Morrell, Pearson, Carter, & Bryant, 1995). In the social sciences, Burke, Shrout, and Bolger (2007) used NLME models to examine individual differences in adjustment to spousal loss; while Grimm and Ram (2009) investigated the effects of preschool instruction on

---

Dr. Harring is Associate Professor of Measurement, Statistics, and Evaluation in the Department of Human Development and Quantitative Methodology. Email him at harring@umd.edu. Dr. Liu is a Psychometrician at ETS.

academic gain using an individual-specific logistic growth model. There are many more examples across diverse research domains.

These applications share several common features. First, mean response for a particular individual is thought to follow a *scientifically-relevant* nonlinear function which characterizes intra-individual behavior in terms of meaningful parameters directly related to the underlying change process. Second, individuals' regression coefficients, in turn, are often formulated to be functions of fixed effects (parameters common to all individuals in the population), covariates (often treatment condition or other individual-level attributes), and individual-specific random effects (parameters representing individual variation). The distribution of random effects captures random variation of the parameters in the population of individuals and is frequently assumed to be multivariate normal.

Although the benefits of incorporating random effects into this framework are undeniable, for a NLME model there is one major drawback. Unlike its linear counterpart (the linear mixed effects model, [Laird & Ware, 1982](#)), one liability is that estimation of model parameters is no longer straightforward. The conditional (on the random effects) mean of the response for an individual depends on the random effects in a nonlinear fashion. This nonlinear dependence requires multidimensional integration over the random effects distribution to derive the needed marginal distribution of the data from which inferences can be made. This integral is almost always intractable having no closed form solution.

Several methods were proposed to overcome this problem. Davidian and Giltinan ([1993](#)) summarized these methods and classified them into four main categories: (1) methods based on individual estimates, (2) methods based on approximating the likelihood through linearizing the nonlinear function, (3) methods based on the exact likelihood which tackle the multidimensional integration directly, and (4) a Bayesian approach which uses both the likelihood based on the data and prior information about model parameters.

The methodological literature has suggested that these methods may not perform equally well under non-ideal data-analytic situations often encountered in practice, including, but not limited to, violation of distributional assumptions, existence of missing data, and small sample sizes. Although a few modest simulation studies were conducted wherein a small subset of these methods were compared for estimating parameters in NLME models, the primary objective of this study was to do a more comprehensive investigation of a broader set of methods across data analytic conditions found in practice presumed to directly impact the estimation methods themselves.

The Nonlinear Mixed-Effects Model

The basic version of the model is considered, although elaborations are possible (see, e.g., Davidian & Giltinan, 2003; Vonesh & Chinchilli, 1997). Following Davidian and Giltinan (1995), the formulation of the nonlinear mixed-effects model for a typical individual selected from the population can be specified in the general form as,

$$\mathbf{y}_i = \mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta}_i) + \mathbf{e}_i, \quad \mathbf{e}_i | \boldsymbol{\beta}_i : [\mathbf{0}, \boldsymbol{\Lambda}_i(\boldsymbol{\lambda})] \quad (1)$$

$$\boldsymbol{\beta}_i = \mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i), \quad \mathbf{b}_i : [\mathbf{0}, \boldsymbol{\Phi}], \quad (2)$$

where $\mathbf{y}'_i = (y_{i1}, \dots, y_{in_i})$ is a $n_i \times 1$ vector of responses, y_{ij} , for the i^{th} individual, $i = 1, K, N$, at times t_{ij} , $j = 1, K, n_i$. Note that the subscript, n_i , on the response implies that the number of measurements and/or the occasions of measurement could vary by individual. Unbalanced data-gathering designs, planned missingness, or data that are missing at random can all be handled by the NLME model in a straightforward fashion. $\mathbf{f}_i(\mathbf{x}_i, \boldsymbol{\beta}_i)$ is an $n_i \times 1$ vector of nonlinear functions with j^{th} element $f(x_{ij}, \boldsymbol{\beta}_i)$, where f is a nonlinear function governing within-individual behavior and is dependent on individual-specific regression parameters $\boldsymbol{\beta}_i$ ($p \times 1$), and x_{ij} contains t_{ij} and other covariates specific to individual i . The $n_i \times 1$ vector of regression residuals, $\mathbf{e}_i$, reflects uncertainty in the response of the i^{th} individual and is assumed to satisfy $E(\mathbf{e}_i | \boldsymbol{\beta}_i) = \mathbf{0}$ for all i . Given the individual coefficients, $\mathbf{y}_i$ has covariance structure $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda})$ which is of dimension $n_i \times n_i$ with $q \times 1$ parameters, $\boldsymbol{\lambda}$, common to all subjects. While many different structures for $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda})$ are possible that reflect various data nuances, when coupled with the random effects covariance structure typically takes on a simple structure such as $\boldsymbol{\Lambda}_i(\boldsymbol{\lambda}) = \sigma^2 \mathbf{I}_{n_i}$. This structure will be used in the forthcoming Monte Carlo simulation.

In the model in Equation 1, variation occurring between individuals is captured through individual-specific parameters, $\boldsymbol{\beta}_i$. Dependence of $\boldsymbol{\beta}_i$ on individual-level covariates $\mathbf{z}_i$ is modeled through $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i)$, a p – dimensional function depending on a $r \times 1$ vector of population parameters $\boldsymbol{\beta}$ and a $k \times 1$ vector of unobservable random effects $\mathbf{b}_i$, associated with individual i . Here, function $\mathbf{g}(\cdot)$ characterizes how elements of $\boldsymbol{\beta}_i$ vary among subjects, due in part to the systematic association with individual attributes, $\mathbf{z}_i$, and unexplained variation in the population captured through $\mathbf{b}_i$. Specifications of $\mathbf{g}(\cdot)$ can be complicated

(see, e.g., Cudeck & Harring, 2007), but at least initially, $\mathbf{g}(\cdot)$ is typically specified as the sum of fixed and random effects such that, $\mathbf{g}(\mathbf{z}_i, \boldsymbol{\beta}, \mathbf{b}_i) = \boldsymbol{\beta} + \mathbf{b}_i$. The variability of the random effects is captured through the $k \times k$ symmetric covariance matrix, $\boldsymbol{\Phi}$. The conventional assumption of normality of the random effects is routinely adopted, but as Hartford and Davidian (2000) state, “simply may be inappropriate.” Numerous scenarios are possible. It may be, for example, that the distribution of the random effects $\mathbf{b}_i$ is skewed or not unimodal. In the latter case, this situation might arise if an important covariate is left out of the model with the resulting systematic variation that would have been attributed to the covariate relegated to the variation in $\mathbf{b}_i$. Consequently, a bimodal or multimodal distribution may be evident, which would not be well-approximated by a normal distribution. In other settings, the distribution of any of the k random effects (b_{ki}) may be symmetric but may be influenced by more cases in the tails of the distribution than would be expected under normality. This might occur because the sample does not accurately reflect the target population and too many individuals in the sample are on the fringe of the distribution resulting in a heavier-tailed distribution with greater dispersion than would be expected otherwise.

A variant of an exponential function will be used in the Monte Carlo simulation. In the social and behavioral sciences, variants of exponential functions are regularly used to summarize the change processes for many phenomena including the learning of a task (see, e.g., Blozis, 2004; Browne, 1993; Meredith & Tisak, 1990), development of language acquisition (Burchinal & Appelbaum, 1991), and growth characteristics (Browne, 1993). Let the individual-specific function, f , characterize the development on a learning task, for example, be an exponential function of the form

$$f(x_{ij}, \boldsymbol{\beta}_i) = \beta_{2i} - (\beta_{2i} - \beta_{1i}) \exp(-\beta_{3i} t_{ij}), \quad (3)$$

which at time t_{ij} for individual i , may provide a suitable summary for intra-individual task performance. The parameters of the model correspond to interesting features of the change process. In Equation 3, β_{1i} represents initial performance when $t_{ij} = 0$, β_{2i} denotes the potential performance at later trials (i.e., $f(t_{ij}) \rightarrow \beta_{2i}$ as $t_{ij} \rightarrow \infty$), and β_{3i} governs the rate of change from initial to potential performance.

Estimation Methods

Much methodological work has been done in recent years for fitting NLME models. The need to derive different approaches may be appreciated by inspection of the form of the marginal distribution of $\mathbf{y}_i$ implied by [Equations 1](#) and [2](#). Denote the conditional density of $\mathbf{y}_i$ given $\mathbf{b}_i$ as $p(\mathbf{y}_i | \mathbf{b}_i)$ and the density of $\mathbf{b}_i$ be denoted as $p(\mathbf{b}_i)$, then the marginal distribution of $\mathbf{y}_i$ is given by

$$p(\mathbf{y}_i) = \int p(\mathbf{y}_i | \mathbf{b}_i) p(\mathbf{b}_i) d\mathbf{b}_i. \quad (4)$$

Define the vector of unique elements in Φ as

$$\begin{aligned} \boldsymbol{\phi} &= \text{vech}(\Phi) \\ &= (\phi_{11}, \phi_{21}, \dots, \phi_{rr})' \end{aligned}$$

where the $\text{vech}(\cdot)$ operator creates a column vector of a symmetric matrix by stacking the diagonal and lower diagonal elements below one another. Putting all relevant model parameters into vector, $\boldsymbol{\theta} : \boldsymbol{\theta}' = (\boldsymbol{\beta}', \boldsymbol{\lambda}', \boldsymbol{\phi}')$, the maximum likelihood estimates for $\boldsymbol{\theta}$ can be found by maximizing in $\boldsymbol{\theta}$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^N \int p(\mathbf{y}_i | \mathbf{b}_i) p(\mathbf{b}_i) d\mathbf{b}_i. \quad (5)$$

Note that, even if both $p(\mathbf{y}_i | \mathbf{b}_i)$ and $p(\mathbf{b}_i)$ are n_i – and k – dimensional normal densities, respectively, $p(\mathbf{y}_i)$ need not be normal. Furthermore, except in a few special cases, the integral will be analytically intractable. Finding a closed form solution is thwarted because $\mathbf{b}_i$ enters function f in a nonlinear manner. In short, inference based on the likelihood of the observed data will be complicated by an inability to express the likelihood in closed form. Therefore, it is crucial to find alternate ways to handle the integration.

Estimation approaches can be categorized into four main categories: (a) methods based on individual estimates, (b) methods based on approximating the likelihood through linearizing the nonlinear function, (c) methods based on the exact likelihood which tackle the multi-dimensional integration directly, and (d) a Bayesian approach which uses both the likelihood based on the data and prior information about model parameters. A thorough description of the aforementioned methods, including complete derivations, may be found in

Wolfinger and Lin (1997), Pinheiro and Bates (1995), Demidenko (2004), and Skrondal and Rabe-Hesketh (2004). A synopsis of each of the methods can also be found on the first author's website (<http://www.education.umd.edu/EDMS/fac/Harring/webpage.html>).

Software Considerations

A self-generated program written in SAS Interactive Matrix Language (IML) was used in the simulation for parameter estimation using the two-stage method based on individuals' estimates with calls to SAS **MIXED** procedure as warranted. Methods based on linearization use algorithms that are numerically simpler than integration methods. They can be found in popular software packages accessible to practitioners. SAS **NLMIXED** procedure was used, based on the First Order (**FIRO**) option (Wolfinger, 1999) for the first-order linearization method. The algorithm of Lindstrom and Bates (1990) conditional first-order method can be obtained by using the **EBLUP** option in the SAS macro **NLINMIX** (Littell et al., 1996). SAS **NLMIXED** was used to implement and execute the Gaussian-Hermite quadrature method using the **NOAD** argument to facilitate the non-adaptive quadrature. Lastly, the **R2WinBUGS** package (Sturtz, Ligges, & Gelman, 2005) in **R** was used to make calls to **WinBUGS** (Spiegelhalter, Thomas, Best, & Lunn, 2002) to facilitate the Bayesian estimation approach. Sample software code for each of these methods can be found in the Appendix.

Review of Previous Simulation Results

Previous simulation studies come from the statistical literature. A non-exhaustive list includes Davidian and Giltinan (1993); Pinheiro and Bates (1995); Roe et al. (1997); Wolfinger and Lin (1997), Hartford and Davidian (2000); Ge, Bickel, and Rice (2004), and Wu (2004).

Davidian and Giltinan (1993) examined the performance of a semiparametric method based on individual estimates and linearization when data had different structures for both inter- and intra-individual variability. They concluded that performance of both methods depended on the relative magnitude of the inter- and intra-individual variability. Misspecification of the intra-individual covariance structure may lead to deterioration in performance for both methods in terms of parameter bias. These methods performed equally well in estimating fixed effects, however, methods based on individual estimates had better estimation of variance and covariance components.

Pinheiro and Bates (1995) examined the performance of the conditional linearization method (Lindstrom & Bates, 1990), Laplace approximation, Gaussian-Hermite and Adaptive Gaussian-Hermite quadrature methods. Their results suggested that the conditional linearization method had the highest computational efficiency but did not provide the most accurate estimation of parameters in terms of bias. Gaussian-Hermite quadrature only provided accurate estimates for large number of quadrature points which made it, in their opinion, computationally inefficient. They concluded that Laplace approximation and Adaptive Gaussian-Hermite quadrature had the best combination of efficiency and accuracy. Pinheiro and Bates' study assumed all assumptions of nonlinear mixed models were met under intensively sampled data. They did not investigate how these methods would perform under distributional misspecification and data sparseness.

Wolfinger and Lin (1997) examined the first-order linearization method and Laplace's approximation method as they are implemented in the SAS macro NLINMIX and concluded that both methods produced reliable estimates, with Laplace's method slightly outperforming the former at the expense of longer computing times and greater instability of the algorithm.

Hartford and Davidian (2000) investigated the consequences for population inference using first-order linearization and Laplace's method when the distribution for the random effects was misspecified – not following a normal distribution. They encountered serious convergence difficulty using Laplace's method when distributions of random effects were far from normal or the population model was not correctly specified. Nevertheless, Laplace's approximation method was still superior to the first-order expansion in parameter accuracy and relative efficiency of estimation except when the random effects distribution was bimodal.

Very little in the NLME model methodological literature has been devoted to how these different estimation methods react to the existence of missing responses or covariates. Wu (2004) suggested that missing values for some of model covariates may have a deleterious effect on parameter recovery. Wu concluded that when the missing data mechanism is nonignorable, serious bias in the parameter estimates may occur.

There has been no simulation work done on the performance of the Bayesian approach. Table 1 provides a summary of the past simulation studies, the estimation methods that were used, the simulation factors that were manipulated, statistical software that was employed if known, and the major findings and limitations.

Research Questions

Specific research questions we address in this simulation study are:

1. Do differences exist between the five estimation methods in terms of parameter bias of fixed effects, variances of the random effects and residual variance? If so, which manipulated study conditions influence the accuracy of parameter recovery?
2. Do differences exist between the five estimation methods in terms of variability of parameter estimates as measured by parameter estimate variance? If so, which manipulated study conditions influence variability of the parameter estimates?

Simulation Design Overview

There are often numerous decision points in analyses involving NLME models. The choice of which method to use often depends on the analytic situation, hypothesis about covariance structures, software availability, sample size, and so on. In order to study the robustness of the five methods of estimating NLME models to the assumptions of normal random effects, conditional normality of the residuals, $\mathbf{e}_i$, data sparseness, and sample size, we carried out a Monte Carlo simulation in which several factors were varied. The data generation model follows [Equations 1 and 2](#), with the exponential model in [Equation 3](#) as the intra-individual function. Although other nonlinear functions could have fewer parameters, we chose this particular function, in part, because it has three coefficients which make the integration feasible, yet is complex enough to examine time to convergence for methods which tackle the integration directly as well as convergence rates for all methods.

Assume that inter-individual function g , is the sum of fixed and random effects

$$\beta_{pi} = \beta_p + b_{pi} \quad p = 1,2,3$$

This simple model specification was chosen so that, hopefully, model identification and convergence issues would be less likely to confound interpretation of performance. Population values for the regression coefficients are $\beta_1 = 100$, $\beta_2 = 10$, and $\beta_3 = 1$. The covariance matrices describing within- and inter-individual variability in [Equations 1 and 2](#), respectively are given as

Table 1. Summary of past simulation and empirical studies on NLME models

Author(s)	Estimation Method	Study Conditions	Summary of Key Findings
Davidian & Giltinan (2003)	• GTS • Other pooled and un-pooled procedures	Intra-individual variability	• Pooling information about intra-individual variability to obtain correct weighting results in improved efficiency • Pooling had little impact on estimation of parameters in β and Φ
Pinheiro & Bates (1995)	• CFO • Laplace • GHQ • Importance Sampling • AGHQ	• Computational efficiency • Parameter estimate comparison • No simulation study	• CFO provides good approximation and is computationally efficient • GHQ is accurate as number of quadrature points increases resulting in computational inefficiency • AGHQ was as accurate as other methods requiring fewer quadrature points and increased computational efficiency
Wolfinger & Lin (1997)	• FO • Laplace	• Normal random effects distribution and no missing data	• Laplace provided less biased estimates but at greater computational cost and instability in the estimation algorithm
Hartford & Davidian (2000)	• FO • Laplace	• Sampling mechanism • Random effects distribution • Population model misspecification	• Laplace converged to a suitable solution with less frequency when model or random effects distribution was misspecified • Estimates under Laplace were generally less biased than FO • No convergence problems under FO method
Ge, Bickel, & Rice (2004)	• CFO • Spline Approximation	• Model followed that of empirical example regularly found in Pharmacokinetics • Random effects distribution	• Inter-individual variability is small, CFO method is efficient and accurate in terms of parameter bias
Wu (2004)	• Exact method of integration • Approximate method of integration	• Response and covariate missingness • Random effects distribution • Sampling mechanism • Error distributions	• Missing data mechanism is non-ignorable, serious bias in the parameter estimates may occur

$$\Lambda_i = \sigma^2 \mathbf{I}_{n_i} \text{ where } \sigma = 2$$

$$\Phi = \begin{pmatrix} \varphi_{11} & & \\ \varphi_{21} & \varphi_{22} & \\ \varphi_{31} & \varphi_{32} & \varphi_{33} \end{pmatrix} = \begin{pmatrix} 25 & & \\ 3 & 4 & \\ 0.05 & 0.05 & 0.075 \end{pmatrix}$$

The empirical performance of each estimation method is evaluated with respect to bias, precision of estimation, and standard error ratios of the fixed parameters β , Φ , and σ^2 . On the basis of $(\hat{\theta}_b : b = 1, \dots, 500)$ obtained from 500 replications, bias is calculated as the differences between the true population values and the means of the estimates obtained from the 500 replications. The variance of the estimates will be used to get some idea as to the precision with which parameters are estimated across study conditions. The variance is computed for the m^{th} element of parameter vector θ as

$$\text{var}[\hat{\boldsymbol{\theta}}(m)] = 500^{-1} \sum_{b=1}^{500} [\hat{\boldsymbol{\theta}}_b(m) - \bar{\hat{\boldsymbol{\theta}}}(m)]^2$$

where for a particular cell, $\bar{\hat{\boldsymbol{\theta}}}(m)$ is the mean of the estimates across the 500 replications, and $\hat{\boldsymbol{\theta}}(m)$ is the estimate obtained by the approach under consideration.

Sample Size and Sampling Scheme

In many applications using NLME models, the sample size is quite small. In a small simulation study, Pinheiro & Bates (1995) used $N = 10$ as the number. In practice, the sample sizes can of course be larger. The total number of subjects will be manipulated to be either: 50, 100, or 250 representing small, medium and large sample sizes, respectively. These correspond to sample sizes found in previous simulation studies (Hartford & Davidian, 2000) as well as empirical studies (see e.g., Cudeck, 1996).

Generated data had a maximum of $n_i = 8$ time points $t_{ij} = 0, \dots, 7$. For all cases, the intra-individual sampling scheme had five total conditions. Data contained either (i) no missingness ($n_i = 8$), (ii) 10% missing, or (iii) 20% missing. Because attrition and drop out seem to occur with some frequency in empirical studies, the missingness was implemented in two ways: (a) deleting the percentage of data for the corresponding time points at the end of the study, and (b) randomly selecting which times would be deleted using the *sample* function in R. R (R Core Team, 2014) was used as the data generation software. The *sample* function in R allows elements from a larger set of elements to be chosen at random.

Data were prohibited at the first time point to be deleted as we felt this was unrealistic in terms of practical data collection protocol – although each of the estimation methods could handle this nuance in a straightforward fashion.

Violation of Normality on Random Effects and Error Distributions

Several different distributions for $\mathbf{b}_i$ were used to generate random effects,

- N. A normal distribution, $\mathbf{b}_i : N(\mathbf{0}, \boldsymbol{\Phi})$

- NN. A non-normal distribution with skew = 2 and kurtosis = 7. The non-normal condition was implemented using the procedure outlined in Headrick and Sawilowsky (1999).
- M1. A mildly contaminated normal distribution, $\mathbf{b}_i : (1 - \pi)N(\mathbf{0}, \mathbf{\Phi}) + \pi N(\mathbf{0}, \mathbf{\Phi}^*)$, with contamination fraction $\pi = 0.05$ and $\mathbf{\Phi}^*$ chosen as described below.
- M2. A moderately contaminated normal distribution, $\mathbf{b}_i : (1 - \pi)N(\mathbf{0}, \mathbf{\Phi}) + \pi N(\mathbf{0}, \mathbf{\Phi}^*)$, with contamination fraction $\pi = 0.10$ and $\mathbf{\Phi}^*$ chosen as described below.

Distribution N denotes the case where the usual assumption of normality on the random effects is applicable. Distribution NN represents a situation where the true distribution of the random effects is positively skewed and heavy-tailed than expected from a normal distribution but with the same variability in the population. Distributions M1 and M2 are meant to characterize the

$$\mathbf{\Phi}^* = \begin{pmatrix} 40 & & \\ 4.5 & 5.5 & \\ 0.07 & 0.07 & 0.095 \end{pmatrix}$$

chosen so that variability is larger but the correlation between effects is approximately the same as those in $\mathbf{\Phi}$. Conceptually, this represents the situation where the apparent inter-individual variation is greater than that in the target population of interest attributable to errors in sampling.

Two distributions for the intra-individual errors, $\mathbf{e}_i$ were used to generate the regression errors

- NE. A normal distribution, $\mathbf{e}_i : N(\mathbf{0}, \sigma^2 \mathbf{I}_{n_i})$
- NNE. A non-normal condition with skew = 3 and kurtosis = 21, respectively. Similarly to the random effects generation, the non-normal condition was implemented using the procedure outlined in Headrick and Sawilowsky (1999).

NMLE MODELS

Distribution NE represents the typical specification of a normal distribution with a simple independence structure. Distribution NNE represents a situation where the true distribution of the regression errors is positively skewed and heavy-tailed than expected from a normal distribution but with the same variability in the population.

A simulation scenario thus consisted of a particular choice of random effects distribution, choice of intra-individual error distribution, sampling scheme, and sample size. The full factorial of $4 \times 2 \times 5 \times 3 = 120$ possible combinations was investigated, where for each scenario, 500 Monte Carlo data sets were generated. For each data set in each scenario, fitting was carried out using each of the five estimation methods as described above. A summary of the manipulated conditions can be found in [Table 2](#).

Table 2. Simulation conditions and levels

Manipulated Condition	# Levels	Levels
Sample Size	3	50, 100, 250
Random Effects Distribution	4	N, NN, M1, M2
Error Distribution	2	NE, NNE
Missingness	5	C, E-10, E-20, R-10, R-20

Note: Levels of the random effects distribution (N = normal, NN = non-normal, M1 = Contaminated 5%, M2 = Contaminated 10%). Levels of the error distribution (NE = normal, NNE = non-normal). Levels of missingness (C = complete cases, E-10 = 10% missing at the end, E-20 = 20% missing at the end, R-10 = 10% randomly missing, R-20 = 20% randomly missing)

Results

The simulations were conducted on several different platforms. The majority of the simulations were completed in a Windows environment on Dell Latitude and Dell Vostro workstations with duo-core processors. Consistency of results was examined across platforms to ensure that conclusions were the results of properties of the methods rather than numerical irregularities. Considering the simulation design, there were 120 fully-crossed conditions for each estimation method, and 500 data sets per scenario. As is often the case in fitting nonlinear mixed effects models by any estimation method, there were some convergence issues and other numerical problems. When numerical problems were encountered, the replicate was repeated with efforts to identify and correct the problem. Despite these efforts several nonconvergent data sets were still present. These trials were categorized as nonconvergent.

In all of the simulation trials, starting values were taken as the true values generating the data to allow the greatest possibility of automation of this large number of simulations. Of course, even in the most optimal condition combinations it may happen that universal convergence can never be achieved. This may be due to poor starting values, practical lack of identifiability with the specific available data, or other unknown factors. Several sets of starting values can be tried to address the first of these issues. However, because of the large number of replications, only limited attempts were made to emulate this “real” practice for initially nonconvergent data sets, which unfortunately did not improve the rate of convergence. The number of data sets (out of 500) for which satisfactory convergence was not achieved for each condition combination and estimation method are shown in [Table 3](#).

There was no convergence problems encountered with the FO (First-order linearization), GHQ (Gaussian-Hermite quadrature), and BAY (Bayesian) methods, although a substantial amount of time was spent preliminarily to examine these methods under worst-case scenario conditions that were thought to influence the successful estimation of the model (i.e., number of quadrature points for the GHQ method, sensitivity of results and convergence to different prior distribution of the parameters for the Bayesian analysis, etc.). The FO method which linearizes the nonlinear function, making it the least computationally intensive method, exhibited no convergence problems what so ever. This is not to say that problems did not occur with these other methods.

The GHQ and FO methods, for example, did not demonstrate lack of convergence based on the default convergence criteria and settings in SAS PROC NLMIXED. Some strange behavior was noticed for several replicate data sets in the Bayesian analysis for the variance components of the model. The reasons behind the odd estimates appears to be that the Bayesian approach is quite sensitive to departures from the assumptions dictated by the prior and data distributions. That is, sensible estimates are not guaranteed for variance-covariance parameters using Bayesian estimation when the underlying distribution is far from the distributions that are presumed in the model set up.

Both the CFO (Conditional first-order) and GTS (Global two-stage) methods showed varying amounts of convergence issues although the number overall was not that significant. It should be noted that unlike the *nlme()* procedure in R, which uses the profiled loglikelihood to stabilize the optimization

NMLE MODELS

Table 3. Rate of nonconvergence out of 500 trials for each distribution, sample size, missingness across estimation method.

SS	ED	Meth	N					NN					M1					M2				
			C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10	C	F-5	F-10	R-5	R-10
50	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	1	0	1	2	0	0	0	1	1	0	1	0	1	1	0	0	0	0
		GTS	0	0	0	2	2	0	0	2	4	4	0	0	2	3	2	0	0	0	3	2
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	1	1	2	1	0	0	0	0	5	0	2	0	2	2	3	0	0
		GTS	3	0	0	4	7	0	0	0	3	3	0	2	0	2	3	2	0	0	3	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
100	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	3	0	0	0	0	3	2	0	0	0	4	2	0	0	0	2	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	2	3	0	0	0	1	0	0	0	0	1	4	0	0	0	0	0
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
250	NE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	2	4	0	0	0	0	3	0	0	0	1	5	0	0	0	0	4
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	NNE	FO	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GHQ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		CFO	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		GTS	0	0	0	0	2	0	2	0	4	3	0	0	0	1	1	0	0	0	2	3
		BAY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: 50, 100, and 250. Missingness levels: C = Complete, E-10 = 10% missing at the end, E-20 = 20% missing at the end, R-10 = 10% randomly missing, R-20 = 20% randomly missing.

algorithm, the *nlinmix* macro in SAS, which was used to estimate the CFO method, does not use profiling. This appears to have some bearing on the stability of the algorithm to estimate parameters under non-ideal conditions. The GTS method uses both nonlinear least squares estimation (which is not affected by distributional assumptions) and PROC MIXED in SAS, which assumes normality in the random effects as well as the data distribution, and therefore could be

susceptible to convergence issues. Surprisingly, this method showed the greatest number of nonconvergent cases among the competitors.

Time to convergence was not an issue for either linearization method (i.e., FO or CFO) as both converged quickly for each replicate with average convergence time of 1.02 and 6.24 seconds, respectfully across all study conditions. Computational speed notwithstanding, time to convergence for these two methods increased as the sample size increased and with random effects distributions that departed from normality. The GTS method was slower to convergence than expected with an average replicate time to convergence of 55 seconds (range of 12.7 seconds under sample size of 50, no missing data, and normal distributions compared with 150.4 seconds per replicate under the most severe study conditions). This may be due to the stage 2 computation using PROC MIXED which utilizes the individual estimates in stage 1 iteratively to compute the variance components of the model. Surprisingly, the GHQ method was faster than expected overall (average time to convergence of 2 minutes per replicate), but suffered a lack of computational speed as the sample size increased and random effects distributions departed from normality. Under these severe conditions, the GHQ method took over 5 minutes to converge. Due to the preliminary investigative analyses, time to convergence for the BAY method was as expected with an average time to convergence of 75 seconds.

ANOVA and Classification Trees

Because of the large number of cells in the design coupled with the numerous parameters and outcomes to evaluate, it is instructive, if not necessary, to use quantitative procedures like analysis of variance (ANOVA) or classification trees as an initial filter of the results – to inform where real effects and “interesting” results occur. Factorial ANOVA was performed on each outcome variable (i.e., bias and parameter estimate variance) for each of the model parameters in $-\beta$, Φ , and σ^2 modeling only main effects as well as two- and three-way interactions. Partial eta-squared, defined as the proportion of total variation attributable to the factor, excluding other factors from the total non-error variation (Pierce, Block & Aguinis, 2004), was used as the arbitrator in deciding which effects to examine more closely, using Cohen’s (1988) heuristic value of (0.14 – large effect) as the cut point.

In conjunction with the ANOVA results, classification trees (Breiman, Friedman, Olshen, & Stone, 1984) were used to aid in determining which factors were most related to each of the outcomes while at the same time establishing

NMLE MODELS

which levels were different from one another. The Chi-squared Automatic Interaction Detection (CHAID) method of constructing each tree (as implemented in SPSS version 20) is an exploratory tool that chooses the independent variable (factor) that has the strongest relation with the dependent variable. Categories of each factor are subsequently merged if they are not significantly different with respect to the dependent variable and the procedure stops when factors (independent variables) no longer affect the outcome. For illustrative purposes, the classification tree for the bias of the estimate of β_2 is shown in Figure 1 below.

Figure 1. Classification tree for bias in β_2 .

The first set of boxes below the initial node represents the method factor as being most related to differences in bias; and the procedure has determined that each method has mean bias that is statistically different from one another with the BAY method showing the least average parameter bias (0.005); the GTS, CFO, and GHQ methods showing comparable values (−0.037, −0.041, and −0.050, respectively); and the FO clearly exhibiting larger average bias than its competitors (0.808). For the FO method it appears that the random effects

distributions which were non-normal and more severely contaminated as a mixture did not seem to have impacted the bias as much as the other distributional conditions. Evidently, no other factors contributed to delineating bias of β_2 further. Nodes representing factors that appear at subsequent levels in this hierarchical structure can be thought of as a type of interaction between itself and the node (or factor) above it. This interaction, however, is specific to particular levels of the factors involved. The entire set of ANOVA and classification tree results as well as tabulated mean bias and variance estimates can be found at the first author's website (<http://www.education.umd.edu/EDMS/fac/Harring/webpage.html>).

Table 4 and Table 5 summarize the results from the ANOVA and classification tree procedures. The ANOVA results for bias in the fixed regression coefficients are displayed in Table 4, which includes the variance components of the random effects associated with these fixed effects, and residual variance. The classification tree results corresponding to the parameters in Table 4 are compiled in Table 5.

Table 4. Main effects, two- and three-way interaction results from a factorial ANOVA for bias of parameters in β , Φ , and σ^2 .

	Bias β_1	Bias β_2	Bias β_3	Bias φ_{11}	Bias φ_{22}	Bias φ_{33}	Bias σ^2
Factor	M	M	M		M	R	
Combi-	R	R	M*R		R	M*R*MI	
nations	M*R	MI			MI	M*MI*S	
		S			S		
		M*MI			M*MI		
		M*R			M*R		
		M*S			M*S		
		M*MI*S			M*R*MI		
		R*MI*S			M*MI*S		
					M*R*S		

Note: M = Method, R = Random Effects Distribution, E = Error Distribution, MI = Missingness, S = Sample Size. The symbol '*' represents the interaction between effects present. To be included, the partial eta-squared for each effect was larger than 0.14 and the effect was significant at the 0.05 level.

Parameter Bias

No main effect or interaction effect was found for the bias in intercept or residual variance, φ_{11} and σ^2 , respectively. Clearly, there were differences in bias across the five methods (M) for each of the regression coefficients; however, method of estimation only influenced the variance of β_{2i} among the variance parameters. This result coincides with the first column (node 1) in Table 5 from the classification tree analysis. Overall, the mean bias values for β_1 , β_2 , and β_3 were

NMLE MODELS

negligible (−0.038, 0.137, −0.016), yet there were differences between the methods. For all regression parameters the FO method showed the greatest bias with the GHQ, CFO, and GTS methods producing less bias estimates. The BAY method constantly generated the least biased estimates (by a factor of 10) compared to the other methods excluding FO. From Table 5, it is clear that for β_1 and β_2 , the random effects distribution significantly impacted the FO method with

Table 5. Results from a classification tree analysis for bias of parameters in β , Φ , and σ^2 .

	Nodes		
	Level 1	Level 2	Level 3
Bias β_1	FO	N/M1 NN/M2	
	GHQ		
	CFO		
	GTS/ BAY	100 50/250	
Bias β_2	FO	N/M1 NN/M2	
	GHQ		
	CFO		
	GTS		
	BAY		
Bias β_3	FO		
	GHQ		
	CFO		
	GTS		
	BAY		
Bias ϕ_{11}	-	-	-
Bias ϕ_{22}	FO	N/M1 NN/M2	
	GHQ	N/NN M1/M2	
	CFO/ GTS	M2 M1 N NN	
	BAY		
Bias ϕ_{33}	N/M1/M2	CFO FO/GHQ/GTS/BAY	
	NN	FO/GTS GHQ/CFO/BAY	
Bias σ^2	-	-	-

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: N = 50, N = 100, and N = 250. The symbol ‘/’ represents levels that are considered the same while levels on different lines are different. Levels are listed from top to bottom in order of magnitude of the bias (greatest to least).

the non-normal and more contaminated mixture distribution producing less bias estimates than the other distributions.

As for the parameters in Φ and σ^2 , in terms of bias, the estimation method, random effects distribution, and combinations of missingness and sample size were consequential. Also, as can be seen in Table 4, the error distribution factor did not influence bias of any parameter in the model including σ^2 . Interestingly, no condition had an effect on the bias of φ_{11} (the variance for β_1), but many conditions, including the amount of missingness, impacted parameter bias for φ_{22} (1.095 overall) with the GHQ method producing less biased estimates on average than the other methods (-0.015). Bias in φ_{33} was negligible (-0.003 overall) even though there were statistical differences across combinations of random effects distributions and methods.

Parameter Variance

In addition to evaluating the accuracy in terms of bias with which these methods produce parameter estimates, precision of estimation is also an important consideration. Table 6 and Table 7 display the summary of results of the factorial ANOVA and classification tree analyses for the variability outcome measure.

Expectedly, sample size was a primary factor in explaining differences in estimate variance with parameter variance decreasing as sample size increased from $N = 50$ to $N = 250$ (0.656 to 0.167). This pattern was evident for all the parameters in which factors impacted variance magnitude. For regression parameters, β_2 and β_3 , precision was also impacted by method and random effects distribution with the GTS and BAY methods producing slightly smaller variance than GHQ with larger discrepancies found in the CFO and FO methods. The ANOVA results coincide with the classification tree results remarkably well, although with slightly different interaction effects. The only variance parameter that showed difference in precision across study conditions was φ_{22} . For this parameter, method seemed to have the most impact with the GHQ and BAY methods producing estimates with the greatest precision (1.43) followed by the CFO and GTS methods (7.72), and lastly the FO method (97.34). When the random effects distribution factor influenced precision, the non-normal distribution frequently produced more precise estimates (less variability) than either of the mixture distributions or normal distribution condition.

NMLE MODELS

Table 6. Main effects, two- and three-way interaction results from a factorial ANOVA for variance of parameter estimates in β , Φ , and σ^2 .

	Var β_1	Var β_2	Var β_3	Var ϕ_{11}	Var ϕ_{22}	Var ϕ_{33}	Var σ^2
Factor Combinations	M	M	M		M		
	R	R	R		R		
	S	S	S		S		
	M*R	M*R			M*R		
	M*S	M*S			M*S		
	R*S	R*S			R*S		
	M*R*S	M*R*S			M*R*S		

Note: M = Method, R = Random Effects Distribution, E = Error Distribution, MI = Missingness, S = Sample Size. The symbol '*' represents the interaction between effects present. To be included, the partial eta-squared for each effect was larger than 0.14 and the effect was significant at the 0.05 level.

Table 7.

	Nodes		
	Level 1	Level 2	Level 3
Var β_1	50	M2 NN/M1 N	
	100	M2 NN/M1 N	
	250		
Var β_2	50	FO/GHQ/CFO GTS/BAY	
	100	FO/CFO GHQ/GTS/BAY	
	250	N/M1/M2 NN	FO/BAY GHQ/CFO/GTS
Var β_3	50	N/M1/M2 NN	FO/GHQ CFO/GTS/BAY
	100	N/M1/M2 NN	FO/GHQ CFO/GTS/BAY
	250	FO/GHQ/GTS CFO/BAY	
Var ϕ_{11}	-	-	-
Var ϕ_{22}	FO GHQ/BAY	50 100 250	
	CFO/GTS	N/M1/M2 NN	50/250 100
Var ϕ_{33}	-	-	-
Var σ^2	-	-	-

Note: Estimation methods: FO = First-Order, GHQ = Gaussian Hermite Quadrature, CFO = Conditional First-Order, GTS = Global Two-Stage, BAY = Bayesian. Random effects distribution levels : N = Normal, NN = Nonnormal, M1 = Contamination 5%, M2 = Contamination 10%. Error distribution levels : NE = Normal, NNE = Nonnormal. Sample size levels: 50, 100, and 250. The symbol '/' represents levels that are considered the same while levels on different lines are different.

Results from large simulation studies are often hard to digest simply by examining tables of values and trying to extract important trends and patterns. The following are the main conclusions from this simulation:

1. Data missingness and error variance distributions seemed to have little if any effect on parameter recovery or estimation precision across the five estimation methods – at least at the levels we investigated.
2. Although the quickest method to converge to a solution and the method least sensitive to starting values, the first-order (FO) linearization method showed the greatest bias across both fixed effects and variance/covariance parameters compared to its competitors.
3. For the other four methods, the GHQ and BAY methods produced the least biased fixed effects although four were comparable for the linear effects.
4. Although slowest time to convergence, the GHQ and BAY methods produced the least biased estimates of the parameters in Φ , while the CFO and GTS methods produced the least biased residual variance. Bias was greatest in these estimates when the sample size was small and/or the random effects distribution was non-normal.
5. Fixed effects were estimated more precisely by the GHQ and BAY methods. For these parameters, precision was affected most by small sample size and non-normal and mixture random effects distributions.
6. Again, the GHQ and BAY methods produced more precise estimates of the variance components of Φ . Expectedly, sample size was also a significant factor variability of the estimates decreasing as the sample size increases.
7. Fixed parameters estimates based on the CFO, BAY, and GTS are fairly robust to mild deviations from normality of both the random effects and error distributions even though these methods sometimes had convergence problems.

NMLE MODELS

These results point to the following recommendations:

1. The FO approach is not recommended for nonlinear mixed effects models as it is the least accurate method for fixed parameter and variance components estimates.
2. The GHQ and BAY methods appear to produce the least biased parameter estimates with the GTS and CFO methods showing comparable results. The GTS, CFO, and BAY methods were more robust to modest departures from normality of the random effects distribution. Thus when the random effects distribution is approximately normal and the sample sizes small to modest, then the GHQ or BAY estimation methods are recommended. For larger sample sizes and deviations from normality, the CFO or the BAY methods are recommended.

The efficacy of the Bayesian approach should be investigated on its own merits and not necessarily compared to likelihood-based methods for estimating nonlinear mixed effects models. This stems from having set up the simulation somewhat unfairly. Apart from the philosophical differences that exist between frequentist and Bayesian approaches, the obvious advantages that the Bayesian framework offers was not exploited. For example, as was previously mentioned, in a Bayesian approach prior knowledge about model parameters including their distributional assumptions can be incorporated into the model formulation. In this simulation, non-informative conjugate priors were used, which put the preponderance of weight in estimating the posterior distribution on the data (or the likelihood). It would be expected that the Bayesian method under this scenario to behave very similarly to the marginal maximum likelihood method, which in this set of simulations it did so unsurprisingly. Further exploration into the methodological underpinnings and extensions of the Bayesian approach that were not investigated here are warranted.

The results of a Monte Carlo simulation study undertaken to gain insight into the consequences of violation of distributional assumptions, sample size, and data sparseness underlying five popular approximations used in fitting nonlinear mixed effects models. Although it is not appropriate to draw general conclusions from a single simulation study, the findings are suggestive and highlight several interesting features that may be worthy of future investigation. It appears that estimation of fixed regression parameters based on the CFO, BAY, and GTS – and to a lesser extent the GHQ approximation – methods is fairly robust to mild

deviations from normality of both the random effects and error distributions, although the GTS method did show difficulty in achieving convergence in a small number of replications. Overall, the FO method showed greater bias than the other methods for the fixed parameters and even more so for the variance components of the model. While it has the least computational burden of any of the methods, it is least accurate and therefore its usefulness in practice is questionable.

Of course, a single simulation cannot possibly examine all of the interesting facets of a model – even if the facility to carry out the computations was limitless. The same could be said of the levels within the manipulated factors that were investigated. Some rationale was provided for the choices knowing that there are infinitely many levels that ultimately could have been chosen. For example, Hartford and Davidian (2000) examined misspecification of the inter-individual model in Equation 2 looking at the performance of both the likelihood ratio test as well as the Wald test to test a single additive component. The current focus was on the estimation of fixed parameters, most of which (β, Φ) characterize the population. Individual regression coefficients, predicted random effects, were not addressed, even though the NLME model is individual-specific. It is not unreasonable to expect that distributional assumptions or other model misspecifications would have more profound effects. Interestingly, methods of carrying out this prediction are markedly different for each of the estimation methods inspected in this study.

Through methodological advances in estimation algorithms and by the sheer speed of today's computing environments, the number of applications using the NLME model has steadily increased – particularly in the social and behavioral sciences. NLME models are important tools for practitioners interested modeling nonlinear change with functions that have at least one regression parameter that enters the function in a nonlinear manner. Much of the methodological and computational techniques for these models were developed in late 1980s through the early 2000s, although some work in the area still exists (Lai & Shih, 2003; Kuhn & Lavielle, 2005; Wu, 2008). As such, many of the estimation methods and optimization schemes for these models have been implemented in popular commercial software. Still a choice for a particular method is required, and often, that choice is made predicated on the research situation and on the specific software being used not necessarily on the merits of the method's performance under sub-optimal, but realistic, data analytic conditions. Overall the results highlight the inherent difficulty in specifying any type of complex model with latent unobservable components; a problem that suggests that caution is in order

in interpreting both the nature of computational issues and results in the event convergence is achieved.

Acknowledgements

This research was partially funded by the Institute of Education Sciences (R305A130042).

References

- Beal, S. L., & Sheiner, L. B. (1982). Estimating population pharmacokinetics. *CRC Critical Reviews in Biomedical Engineering*, 8, 195-222.
- Blozis, S. A. (2004). Structured latent curve models for the study of change in multivariate repeated measures data. *Psychological Methods*, 9(3), 334-353. doi: [10.1037/1082-989X.9.3.334](https://doi.org/10.1037/1082-989X.9.3.334)
- Blozis, S. A., & Cudeck, R. (1999). Conditionally linear mixed-effects models with latent variable covariates. *Journal of Educational & Behavioral Statistics*, 24(3), 245-270. doi: [10.3102/10769986024003245](https://doi.org/10.3102/10769986024003245)
- Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. New York: Chapman & Hall/CRC.
- Brooks, S. P., & Gelman A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7(4), 434-455. doi: [10.1080/10618600.1998.10474787](https://doi.org/10.1080/10618600.1998.10474787)
- Browne, M. W. (1993). Structured latent curve models. In C. M. Cuadras & C. R. Rao (Eds.), *Multivariate analysis: Future directions 2* (pp. 171-197). Amsterdam: Elsevier Science.
- Burchinal, M., & Appelbaum, M. I. (1991). Estimating individual developmental functions: Methods and their assumptions. *Child Development*, 62(1), 23-43. doi: [10.2307/1130702](https://doi.org/10.2307/1130702)
- Burke, C. T., Shrout, P. E., & Bolger, N. (2007). Individual differences in adjustment to spousal loss: A nonlinear mixed model analysis. *The International Journal of Behavioral Development*, 31(4), 405-415. doi: [10.1177/0165025407077758](https://doi.org/10.1177/0165025407077758)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

- Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, 31(3), 371-403. doi: [10.1207/s15327906mbr3103_6](https://doi.org/10.1207/s15327906mbr3103_6)
- Cudeck, R., & Harring, J. R. (2007). The analysis of nonlinear patterns of change with random coefficient models. *Annual Review of Psychology*, 58, 615-637. doi: [10.1146/annurev.psych.58.110405.085520](https://doi.org/10.1146/annurev.psych.58.110405.085520)
- Davidian, M., & Giltinan, D. M. (1993). Some general estimation methods for nonlinear mixed-effects models. *Journal of Biopharmaceutical Statistics*, 3(1), 23-55. doi: [10.1080/10543409308835047](https://doi.org/10.1080/10543409308835047)
- Davidian, M., & Giltinan, D. M. (1995). *Nonlinear models for repeated measurement data*. London: Chapman & Hall.
- Davidian, M., & Giltinan, D. M. (2003). Nonlinear models for repeated measurements: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8, 387-419. doi: [10.1198/1085711032697](https://doi.org/10.1198/1085711032697)
- Demidenko, E. (2004). *Mixed models*. New York: Wiley
- Ge, Z., Bickel, P. J., & Rice, J. A. (2004). An approximate likelihood approach to nonlinear mixed effects models via spline approximation. *Computational Statistics & Data Analysis*, 46(4), 747-776. doi: [10.1016/j.csda.2003.10.011](https://doi.org/10.1016/j.csda.2003.10.011)
- Grimm, K. J., & Ram, N. (2009). Nonlinear growth models in Mplus and SAS. *Structural Equation Modeling, A Multidisciplinary Journal*, 16(4), 676-701. doi: [10.1080/10705510903206055](https://doi.org/10.1080/10705510903206055)
- Harring, J. R., Cudeck, R., & du Toit, S. H. C. (2006). Fitting partially nonlinear random coefficient models as SEMs. *Multivariate Behavioral Research*, 41, 579-596. doi: [10.1207/s15327906mbr4104_7](https://doi.org/10.1207/s15327906mbr4104_7)
- Hartford, A., & Davidian, M. (2000). Consequences of misspecifying distributional assumptions in nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 34(2), 139-164. doi: [10.1016/S0167-9473\(99\)00076-6](https://doi.org/10.1016/S0167-9473(99)00076-6)
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, 64(1), 25-35. doi: [10.1007/BF02294317](https://doi.org/10.1007/BF02294317)
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, 42(4), 805-820. doi: [10.2307/2530695](https://doi.org/10.2307/2530695)

- Kuhn, E., & Lavielle, M. (2005) Maximum likelihood estimation in nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 49(4), 1020-1038. doi: [10.1016/j.csda.2004.07.002](https://doi.org/10.1016/j.csda.2004.07.002)
- Lai, T. L., & Shih, M. C. (2003). Nonparametric estimation in nonlinear mixed effects models. *Biometrika*, 90(1), 1-13. doi: [10.1093/biomet/90.1.1](https://doi.org/10.1093/biomet/90.1.1)
- Laird, N. M., & Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics*, 38(4), 963-974. doi: [10.2307/2529876](https://doi.org/10.2307/2529876)
- Lindstrom, M. J., & Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3), 673-687. doi: [10.2307/2532087](https://doi.org/10.2307/2532087)
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS® System for Mixed Models*. Cary, NC: SAS Institute Inc.
- Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New Jersey: Springer.
- Martins, C. R., Oulhaj, A. A., de Jager, C. A., & Williams, J. H. (2005). APOE alleles predict the rate of cognitive decline in Alzheimer disease: A nonlinear model. *Neurology*, 65(12), 1888-1893. doi: [10.1212/01.wnl.0000188871.74093.12](https://doi.org/10.1212/01.wnl.0000188871.74093.12)
- Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, 55(1), 107–122. doi [10.1007/BF02294746](https://doi.org/10.1007/BF02294746)
- Paap, R. (2002). What are the advantages of MCMC based inference in latent variable models? *Statistica Neerlandica*, 56(1), 2-22. doi: [10.1111/1467-9574.00060](https://doi.org/10.1111/1467-9574.00060)
- Pierce, C. A., Block, R. A., & Aguinis, H. (2004). Cautionary note on reporting eta-squared values from multifactor ANOVA designs. *Educational & Psychological Measurement*, 64(6), 916–924. doi: [10.1177/0013164404264848](https://doi.org/10.1177/0013164404264848)
- Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the loglikelihood function in the nonlinear mixed effects model. *Journal of Computational and Graphical Statistics*, 4(1), 12-35. Doi: [10.1080/10618600.1995.10474663](https://doi.org/10.1080/10618600.1995.10474663)
- Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer-Verlag.
- R Core Team (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org>.
- Roe, D. J., Vonesh, E. F., Wolfinger, R. D., Mesnil, F., & Mallet, A. (1997). Comparison of population pharmacokinetic modeling methods using simulated data: results from the Population Modeling Workgroup. *Statistics in Medicine*,

16(11), 1241-1262. doi: [10.1002/\(SICI\)1097-0258\(19970615\)16:11<1241::AID-SIM527>3.0.CO;2-C](https://doi.org/10.1002/(SICI)1097-0258(19970615)16:11<1241::AID-SIM527>3.0.CO;2-C)

Segawa, E. (1998, November). Application of hierarchical nonlinear models to children's growth in vocabulary and height. *Dissertation Abstracts International*, 59, Retrieved from EBSCOhost.

Sheiner, L. B., & Beal, S. L. (1980). Evaluation of methods for estimating population pharmacokinetic parameters, I. Michaelis-Menten model: Routine clinical pharmacokinetic data. *Journal of Pharmacokinetics and Biopharmaceutics*, 8(6), 553-571. doi: [10.1007/BF01060053](https://doi.org/10.1007/BF01060053)

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*. London: Chapman & Hall.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2002). *WinBugs User Manual (Version 1.4)*. Cambridge, UK: MRC Biostatistics Unit 26.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1-16. doi: [10.18637/jss.v012.i03](https://doi.org/10.18637/jss.v012.i03)

Vonesh, E. F., & Chinchilli, V. M. (1997). *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. New York: Marcel Dekker.

Wolfinger, R. D. (1993). Laplace's approximation for nonlinear mixed models, *Biometrika*, 80(4), 791-795. doi: [10.1093/biomet/80.4.791](https://doi.org/10.1093/biomet/80.4.791)

Wolfinger, R. D. (1999). *Fitting Nonlinear Mixed Models with the New NLMIXED Procedure, Paper 287* [Technical Report]. Cary, NC: SAS Institute Inc.

Wolfinger, R. D., & Lin, X. (1997). Two Taylor-series approximation methods for nonlinear mixed effects models. *Computational Statistics and Data Analysis*, 25(4), 465-490. doi: [10.1016/S0167-9473\(97\)00012-1](https://doi.org/10.1016/S0167-9473(97)00012-1)

Wu, L. (2004). Nonlinear mixed effects models with nonignorably missing covariates. *Canadian Journal of Statistics*, 32(1), 27-37. doi: [10.2307/3315997](https://doi.org/10.2307/3315997)

Wu, L. (2008). An approximate method for nonlinear mixed-effects models with nonignorably missing covariates. *Statistics & Probability Letters*, 78(4), 384-38. doi: [10.1016/j.spl.2007.07.011](https://doi.org/10.1016/j.spl.2007.07.011)

Appendix A

Data for the simulation was generated in R (V 3.0.1). The following input statements were used to run each of the methods in the various statistical software programs.

First-Order Linearization (FIRO) Using SAS PROC NLMIXED

```
proc nlmixed data=aera method=firo tech=quanew lis=2 lsp=.005 maxfu=5000
maxit=2000;
parms au=100 bu=10 cu=1 sa=25 sb=1, sc=0.075 sab=3 sac=0.05 sbc=0.05 se=4;
a=au+ai;
b=bu+bi;
c=cu+ci;
mod= b-(b-a)*exp(-c*(time-1));
model aera ~ normal(mod,se);
random ai bi ci ~ normal([0,0,0],[sa,sab,sb,sac,sbc,sc]) subject=id;
run;
```

Global Two-Stage (GTS) Using SAS Macro

```
%macro GTS(size);
proc iml;
print &size;

*first stage estimate of individual person parameter;
%do k=1 %to 100;
proc iml;
use aera.aera;
read all;
dat=time||y||id;
uid=t(unique(id));
m=nrow(uid);
n=nrow(id);
p=3;
dati=J(8,3,0);
create indivdat from dati [colname={'time' 'y' 'subj'}] ;
do i=1 to 8;
dati[i,]=dat[i+(&k-1)*8,];
end;

append from dati;
quit;

proc nlin data=indivdat noprint save outest=test ;
parms b1=100, b2=10, b3=1;
model y=b2-(b2-b1)*exp(-b3*(time-1));
output out=nlinout predicted=pred residual=res ;
run;
data par; set test; if _type_ ne "FINAL" then delete; subj=&k; keep subj
_status_ b1 b2 b3; run;
```

```

proc append base=stage1par data=par force;
proc append base=stage1pred data=nlinout force;
%end;
proc iml;
* read in nlin estimated results;
use stage1par;
read all into bols [colname=name];
use stage1pred;
read all;
n=nrow(pred);
m=nrow(bols);
p=3;

*pooled ols estimate of sigma;
sigma=sum(res#res)/(n-m*p);
create var_e from sigma [colname={'error variance'}];
append from sigma;

*get covariance matrix for each bols;
*prepare data for proc mixed analysis;
thisdati=J(3,8,0);
create mixdat from thisdati [colname={'id' 'y' 'x1' 'x2' 'x3' 'z1' 'z2' 'z3'}];
prednew=J(m,8,0);
grd1=J(1,8,0);
grd2=J(1,8,0);
grd3=J(1,8,0);
x={1,2,3,4,5,6,7,8};
do l=1 to 100;
  do j=1 to 8;
    prednew[l,j]=bols[l,2] -(bols[l,2]-bols[l,1])*exp(-bols[l,3]*(x[j]-1));
    grd1[j]=exp(-bols[l,3]*(x[j]-1));
    grd2[j]=1-exp(-bols[l,3]*(x[j]-1));
    grd3[j]=(bols[l,2]-bols[l,1])*(x[j]-1)*(exp(-bols[l,3]*(x[j]-1)));
    grd=t(grd1)||t(grd2)||t(grd3);
  end;
  thisid=J(p,1,1);
  bi=I(3);
  A=sigma*solve(t(grd)* grd,bi);
  chalf=root(solve(A,bi));
  respi=chalf*t(bols [l,1:3]) ;
  thisxi=chalf;
  thisdati=thisid||respi||thisxi||thisxi;
  append from thisdati;
end;
quit;

*final population parameter estimate;
proc mixed data=mixdat method=ml covtest;
  class id;
  model y = x1 x2 x3 / noint solution chisq;
  random z1 z2 z3/ subject=id type=un g gcorr gc;
  parms (25) (3) (1) (0.06) (0.06) (0.075) (1) / eqcons=7;
  ods output solutionf=fixedparms;
  ods output CovParms=covparms;

```

NMLE MODELS

```
run;  
%mend GTS;
```

Conditional First-Order Linearization (CFO) Using SAS Macro NLINMIX

```
%nlinmix(data=dat,  
  model=%str(  
    a=au+ai;  
    b=bu+bi;  
    c=cu+ci;  
    predv= b-(b-a)*exp(-c*(time-1));  
  ),  
  parms=%str(au=100 bu=10 cu=.75),  
  stmts=%str(  
    class id;  
    model pseudo_y = d_au d_bu d_cu / noint notest solution cl;  
    random d_ai d_bi d_ci / type=un subject=id solution;  
  ),  
  expand=eblup  
),  
run;
```

Gaussian-Hermite Quadrature (GHQ) Using SAS PROC NLMIXED

```
proc nlmixed data=aera method=gauss noad tech=quanew lis=2 lsp=.005 maxfu=5000  
maxit=2000 qpoints=20;  
parms au=100 bu=10 cu=1 sa=25 sb=1, sc=0.075 sab=3 sac=0.05 sbc=0.05 se=4;  
a=au+ai;  
b=bu+bi;  
c=cu+ci;  
mod= b-(b-a)*exp(-c*(time-1));  
model aera ~ normal(mod,se);  
random ai bi ci ~ normal([0,0,0],[sa,sab,sb,sac,sbc,sc]) subject=id;  
run;
```

Bayesian (BAY) Using R and WinBUGS

The Bayesian approach used the R2WinBUGS library and bugs function in R. R was utilized as the platform to call WinBUGS and collate results upon convergence of the program. There is a debugging option in the bugs function that allows monitoring of the iteration history and mixing. We used this extensively in the beginning to identify problematic code. The bugs function requires three files to call the WinBUGS program:

```
nlme.sim <- bugs(data, inits, parameters, "C:/ /programs/  
quadwin.txt",  
  n.chains=3, n.iter=9000, n.burnin=7000,  
  bugs.directory="C:/Program Files/WinBUGS14",
```

```
n.thin = 1, debug=T)
```

File 1: Initial Values (init)

```
inits = function(){
  list(mub=c(100,10,1),
       tau=matrix(c(.05,0,0,0,.25,0,0,0,20),nrow=3,byrow=F),
       tauC=.5)
}
```

File 2: Parameters to Monitor (parameters)

```
parameters = c("mub", "sig", "sige")
```

File 3: Model Statement (quadwin.txt)

```
model {
  for (i in 1:K) {
    for (j in 1:n) {
      z[i, j] ~ dnorm(mnb[i, j], tauC)
      mnb[i, j] <- b[i, 2] - ((b[i,2] - b[i,1])*exp(-b[i,3] * x[j]))
    }
    b[i, 1:3] ~ dmnorm(mub[1:3], tau[1:3,1:3])
  }

  mub[1:3] ~ dmnorm(mean[1:3], S2[1:3,1:3])
  tau[1:3, 1:3] ~ dwish(S3[1:3,1:3], 3)
  sigma2[1:3, 1:3] <- inverse(tau[1:3,1:3])
  sig[1,1] <- sigma2[1,1]
  sig[1,2] <- sigma2[2,1]
  sig[2,2] <- sigma2[2,2]
  sig[1,3] <- sigma2[3,1]
  sig[2,3] <- sigma2[3,2]
  sig[3,3] <- sigma2[3,3]
  tauC ~ dgamma(1.0E-3, 1.0E-3)
  sige <- 1 / tauC
}
```

A Spatial Analytical Framework For Examining Road Traffic Crashes

Grace O. Korter
University of Ibadan
Ibadan, Nigeria

A number of different modeling techniques have been used to examine road traffic crashes for analytic and predictive purposes. Map-based spatial analysis is introduced. Applications are given which show the power in a combination of existing exploratory and statistical methods.

Keywords: Box plots, Cartograms, Maps, Regression, Spatial autocorrelation

---

Introduction

Road Traffic Crashes (RTC) data are obtained for observations which are ordered in space or in space and time. In these situations, the observations can be characterized by their absolute location, using a coordinate system, or by their relative location, based on a particular distance metric. In other words, RTC data are organized by spatial units of observation across geographical space in the most general sense.

In many instances where regression analysis is applied to spatial data, various measurement problems, spatial externalities, spatial dependence and other spatial effects which may lead to dependent error terms could arise. As such, the standard econometric modeling techniques become invalid. The main methodological problems stem from the existence of spatial effects in the form of spatial dependence (also known as spatial autocorrelation). For instance, spatial autocorrelation impacts mildly on regression coefficient estimates for geographically distributed non distance based covariates, regardless of whether the statistical distribution of a variable is Normal, Poisson or Binomial. In contrast spatial autocorrelation impacts significantly upon the variance of these variables, with positive autocorrelation inducing variance inflation. Also, this increased

---

*Dr. Korter is a Lecturer at Federal Polytechnic Offa. Email her at:
kortergo@gmail.com.*

variance tends to flatten a bell shaped curve and create over dispersion in Poisson and Binomial variables. Consequently, frequency distributions are distorted when spatial autocorrelation is present.

As positive spatial autocorrelation increases, for a normal curve, the center is squashed and the tails are thickened. Whereas, the shape of a Poisson distribution increasingly resembles that of an exponential distribution, with the frequency of zero and/or near-zero values increasing concomitantly with the appearance of marked outliers. Also, the shape of a Binomial distribution first approaches that of a Uniform distribution and then that of a Sinusoidal distribution.

These metamorphoses are the only way variance can increase for these types of variables. In addition, spatial autocorrelation compromise the descriptive ability of conventional versions of the above probability models, as well as the statistical inferences associated with them.

The majority case of positive spatial autocorrelation results in variance inflation, which for a normal frequency distribution means the central frequencies (i.e., numbers of values near the mean) decrease and the tail frequencies increase - peakedness (i.e., kurtosis) is altered because the curve is flattened. The most common case of positive spatial autocorrelation also results in variance inflation for a Poisson frequency distribution. Here, the chances of smaller and of extremely large values increase at the expense of intermediate values. Again, positive spatial autocorrelation also, results in variance inflation for a binomial frequency distribution, which means the chances of the extreme probabilities increase at the expense of more centrally located intermediate probabilities.

Exploratory data analysis and spatial autoregressive models have brought new power to the analysis of spatial data, capturing effects of redundant locational information contained in these data. Their descriptive abilities better portray observed data and they offer more fully developed nomothetic approach that in turn will support an increasing emphasis on the geographically varying local statistics that fluctuate around their global parameter counterparts (see [Anselin & Griffith \(1988\)](#), [Anselin \(1988\)](#), [Anselin \(1990\)](#), [Griffith \(2009\)](#), [LeSage & Pace \(2009\)](#)).

A great amount of effort has gone into assessing the different factors that contribute to RTCs. Previous studies have utilized a number of different modeling techniques for both analytical and predictive purposes. [Jovanis and Chang \(1986\)](#), [Maher and Mountain \(1988\)](#), [Jones, Jannsen and Mannering \(1991\)](#) and [Roque and Cardoso \(2014\)](#) applied Poisson regression models to study aggregate RTC

data, while, Ye et al. (2009) and Yu and Abdel-Aty (2013) investigated the frequencies of RTC using multivariate Poisson regression model.

Again, the Poisson based models; Poisson-Lognormal, Poisson-Gamma and Poisson-Lognormal with conditional autoregressive priors have been found suitable for RTC studies and these have been widely utilized in literature. Examples include the works of Shanker, Mannering and Barfield (1995), Sheffer and Riet Veld (1997), Milton and Mannering (1998), Abdel-Aty and Radwan (2000), Amoros, Martin and Laumon (2003), Miaou and Lord (2003), Noland and Quddus (2005), Kim, Brunner and Yamashita (2006), Lord and Miranda-Moreno (2008), Quddus (2008) and Wang, Quddus and Ison (2009).

Also, the negative binomial model was used by Jovanis and Chang (1986), Fridstrøm and Ingebrigtsen (1991), Miaou and Lord (2003), Guo, Wang and Abdel Aty (2010), Ahmed et al.(2011) , Yu, Abdel Aty and Ahmed (2013) and Goh, Currie, Sarvi and Logan (2014) to model RTC.

Using the Bayes approach, Aguero-Valverde and Jovanis (2006) and Jiang, Abdel Aty and Alamili (2014) employed random effect Poisson Log-Normal models for crash risk hotspot identification. While, Yu and Abdel-Aty (2013) utilized the random effects logistic regression models to investigate the dynamics of RTC.

Similarly, Jegede (1988), Okamoto and Koshi (1989) and Aderamo (2012) used the multiple regression model to study the incidence of RTC. However, Anastasopoulos, Tarko and Mannering (2008) identified the tobit regression model as an alternative method for studying RTC. In addition, Erdogan (2009) studied RTC using the geographically weighted regression model, while, Abdullah and Zamri (2012) applied the fuzzy regression model using two threshold levels. Taking cognizance of spatial autocorrelation, Levin, Kim, and Nitz (1995a, 1995b) and Korter, Olubusoye and Salisu (2014) modeled RTC using the spatial autoregressive model.

Methodology

Data on number of licensed drivers, registered vehicles and 2011 RTC was obtained from the Federal Road Safety Commission (FRSC). The land area encompassing each state and the Federal Capital Territory (FCT) was obtained, while, the corresponding population was obtained from National Population Commission.

Cartograms

To illustrate the relative sizes of the RTC for the states/FCT within the country, the area of each state/FCT is scaled in proportion to its RTC.

Cartograms are unique representations of geographical space that depict spatial transformations and attributes of geographic objects as the object's area. Examined more closely, the value-by-area mapping technique encodes the mapped data in a simple and efficient manner with no data generalization or loss of detail. Cartograms vary on the degree in which geographic space is changed.

A cartogram is yet another way to highlight extreme values on a map. Here the original spatial units are replaced by circles. The area of the circle is proportional to the value of a selected variable. The circles themselves are aligned as closely as possible to the original location of the matching spatial units by means of a nonlinear optimization routine.

A Dorling cartogram maintains neither shape, topology nor object centroids, though it has proven to be a very effective cartogram method. To create a Dorling cartogram, instead of enlarging or shrinking the objects themselves, the cartographer will replace the objects with a uniform shape, usually a circle, of the appropriate size. Circles are distributed to resemble the original topology. This method suggests that the shapes do not overlap but rather be moved so that the full area of each shape can be seen. The shape and relative location of each state and FCT is retained to as large an extent as possible (see [Dorling, 1996](#); [Anselin 2005](#)).

Box Plots

Box plots are good at portraying extreme values and are especially good at showing differences between distributions. It is a type of graph which is used to show the shape of the distribution, its central value, and variability. The picture produced consists of the most extreme values in the data set (maximum and minimum values), the lower and upper quartiles, and the median.

In descriptive statistics, a box plot is a convenient way of graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram. Outliers may be plotted as individual points.

Box plots display variation in samples of a statistical population without making any assumptions of the underlying statistical distribution: box plots are non-parametric. The spacings between the different parts of the box indicate the

degree of dispersion (spread) and skewness in the data, and show outliers. In addition to the points themselves, they allow one to visually estimate the interquartile range, midhinge, range, mid-range and trimean. Boxplots can be drawn either horizontally or vertically.

Here, the rectangle represents the cumulative distribution of RTC sorted by value. The red bar in the middle corresponds to the median; the dark part shows the interquartile range (going from the 25th percentile to the 75th percentile). The individual observations in the first and fourth quartile are shown as blue dots. The thin line is the hinge which corresponds to the default criterion of 1.5. The hinge criterion determines how extreme observations need to be before they are classified as outliers (see McGill, Tukey & Larsen, 1978; Anselin, 2005).

The Spatial Autoregressive Model Specification

$$y = \rho W(y) + X\beta + \varepsilon \quad (1)$$

where, y is an $N \times 1$ vector of the dependent variable (RTC cases) for all 36 states and the FCT. This is formed by observing the total incidence of RTC for each of the 36 states and the FCT. ρ is the coefficient of the spatial lag. W is a known $N \times N$ spatial weights matrix whose diagonal elements are zero. Two states are said to be contiguous to one another when they share a common border. A spatial weights matrix is defined by using an interaction between the geographical location of each spatial unit and every other location. The neighbourhood contiguity distance between each location and every other location was used.

To develop the spatial weights (contiguity based matrix), the value of 1 was assigned to a state when it shares common border with another, and otherwise the value 0 was assigned. Using the common border, the rook contiguity matrix was used to form a weights matrix. Taking into cognizance the common border and the common vertex border, the queen contiguity matrix was formed. W satisfies the condition that $(I_N - \rho W)$ is non singular for all $|\rho| < 1$, I_N is an identity matrix of dimension N . $W(y)$ is a weighted matrix of $N \times 1$ vector of values for the RTC cases summed over all states.

X is an $N \times K$ matrix of observations on the K explanatory variables, namely population, drivers, area and vehicles. X is assumed to be of full column rank and its elements are assumed to be asymptotically bounded in absolute value. β is a $K \times 1$ vector of K parameters, representing the 4 regression coefficients for this

study. ε is an $N \times 1$ vector of normally distributed random error terms, with mean and constant variance, while, $E(X / \varepsilon) = 0$, strict exogeneity.

The ordinary least squares estimators are obtained from a regression of the X on y and on Wy respectively. Clearly, the maximum likelihood estimate for β is a function of these auxiliary regression coefficients as well as of ρ . Therefore, the estimate for β can be found directly, once the value for ρ has been determined. These steps can be carried out by standardized regression package. This study employed the GeoDa software ([Anselin, 2005](#)).

Predicted Value and Residual Maps

The predicted value and regression residuals are variables to any exploratory function in GeoDa, including mapping. Such maps, referred to as predicted value maps and residual maps are useful for a visual inspection of patterns. The predicted value map can be thought of as a smoothed map, in the sense that random variability, due to factors other than those included in the model has been smoothed out. A residual map may give an indication of systematic over- or under prediction in particular regions, which could be evidence of spatial autocorrelation, although this needs to be assessed more vigorously by means of a hypothesis test.

The residual map does suggest that similarly colored areas tend to be in similar locations, which could indicate positive spatial autocorrelation. Also, it indicates a tendency to over predict (negative residuals) in the outlying areas and a tendency to under predict (positive residuals) in the core, suggesting the possible presence of spatial heterogeneity in the form of spatial regimes ([Anselin, 2005](#)).

Results

From [Figure 1](#), extreme value of RTC was found in Abuja (red circle). The numbers of RTC across the 36 states was normal (green circles). The size of each circle is proportional to the number of RTC in the spatial units (the states).

SPATIAL ANALYTICAL FRAMEWORK

Figure 1. Dorling Cartogram of RTC (2011) in Nigeria

The upper outlier is Abuja. Niger state is the central position of RTC in Nigeria. 50 percent of the states have RTC number greater/lower than the number of RTC (88) that occurred in Niger state. 25 percent of the states had above 222 RTC cases in 2011, while the remaining 75 percent had RTC cases ranging from 16-180. Nine states had between 16-52 RTC cases, these include Akwa Ibom, Anambra, Bayelsa, Borno, Cross River, Ebonyi, Jigawa, Kebbi and Yobe states. In ten other states RTC cases ranged between 54 and 88. This includes Abia, Adamawa, Ekiti, Enugu, Gombe, Katsina, Niger, Plateau, Sokoto and Lagos states. Nine other states, Bauchi, Zamfara, Delta, Edo, Imo, Kaduna, Rivers and Taraba had between 90-180 RTC cases. At least 222 number of RTC cases were recorded in Benue, Kano, Kogi, Nassarawa, Ogun, Ondo, Osun and Oyo states (see Figure 2). Figure 3 allows a virtual inspection of the distributions across the spatial units.

Figure 2. Box plot of RTC (2011) in Nigeria

Figure 3. Box plot Map of RTC (2011) in Nigeria

SPATIAL ANALYTICAL FRAMEWORK

Spatial Autoregressive Model

Adopting the queen contiguity based spatial weights matrix, the spatial autoregressive model was built (Table 1). The number of observations equal 37, number of variables 6, while the degrees of freedom equal 31. The R-squared equal 50%. The mean of the dependent variable (RTC) is 128.78 and the standard deviation equal 104.74. The residual variance (sigma-square) is 5460.34, while, the standard error estimate (standard error regression) is 73.89.

A limited number of diagnostics are provided with the maximum likelihood lag estimation. First is the Breusch-Pagan test for heteroskedasticity in the error terms. The highly insignificant value of 0.88 (p -value 0.93) suggests that heteroskedasticity is not a serious problem. The second test is an alternative to the asymptotic significance test on the spatial autoregressive coefficient; it is not a test on remaining spatial autocorrelation. The value 5.88 (p -value 0.01) confirms strong significance of the spatial autoregressive coefficient.

The estimated ρ is positive and significant, indicating moderate spatial autoregressive dependence in RTC. In other words, RTC tend to be more clustered by states than what would be expected by a random distribution. The number of RTC cases for a given state is affected by the RTC cases of the neighbouring states.

Table 1. Spatial Autoregressive Model Estimates

Variable	Coefficient	Std. Error	z-value	Probability
Spatial Lag	0.45	0.17	2.64	0.01
Constant	1.44	39.29	0.04	0.97
Drivers	0.01	0.02	4.54	0.05
Vehicles	0.65	0.01	0.04	0.96
Area	-0.01	0.01	-0.21	0.83
Population	0.87	0.89	1.08	0.28

The parameter estimates for population, vehicles and area are not significant. This means that each of these exogenous variables does not contribute significantly to the incidence of RTC in Nigeria. However, the sign of the coefficient suggests the following:

Controlling for the spatial lag, population is positively related to the number of RTC occurring within the localities. Indicating that population generate a certain level of RTC. Number of registered vehicles characteristic produced a positive sign. The land area encompassing each state has a negative sign. All other things being equal, states with larger residential populations tend to have

more RTC and states with larger number of registered vehicles tend to have more RTC. Although, increase in the area of administration suggests less RTC.

Spatial units with higher than expected future likelihood of RTC include the FCT-Abuja, Kaduna, Kano, Oyo, Edo and Rivers states (Figure 4). States that seem to have high spatial dependence and where the RTC tend to be over predicted include Zamfara, Nassarawa, Benue, Kogi, Ondo and Osun states (Figure 5).

Figure 4. Predicted Value Map

Figure 5. Residual Map

Conclusion

In this paper, a combination of statistics and geographic information system is proposed. The applications have shown the power in the descriptive ability in statistical analysis. While, the proposed model uses a combination of statistics and geo visual representation of parameter estimates, tests for spatial autocorrelation is another option to increase emphasis on spatial dependencies. However, the proposed model has the advantage of giving a better understanding of parameter estimates.

References

- Abdel-Aty, M. A. & Radwan, A. E. (2000). Modelling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5), 633-642. doi:[10.1016/S0001-4575\(99\)00094-9](https://doi.org/10.1016/S0001-4575(99)00094-9)
- Abdullah L. & Zamri N. (2012). Road accident models with two threshold levels of fuzzy linear regression. *Journal of Emerging Trends in Computing and Information Sciences*, 3(2), 225-230.
- Aderamo A. J. (2012). Spatial pattern of road traffic accident casualties in Nigeria. *Mediterranean Journal of Social Sciences*, 3(2), 61-72. doi:[10.5901/mjss.2012.v3n2.61](https://doi.org/10.5901/mjss.2012.v3n2.61)
- Aguero-Valverde, J. & Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38(3), 618-625. doi:[10.1016/j.aap.2005.12.006](https://doi.org/10.1016/j.aap.2005.12.006)
- Ahmed, M., Huang, H., Abdel-Aty, M. & Guevara, B. (2011). Exploring a Bayesian hierarchical approach for developing safety performance functions for a mountainous freeway. *Accident Analysis & Prevention*, 43(4), 1581-1589. doi:[10.1016/j.aap.2011.03.021](https://doi.org/10.1016/j.aap.2011.03.021)
- Amoros, E., Martin, J. L. & Laumon, B. (2003). Comparison of road crashes incidence and severity between some French countries. *Accident Analysis & Prevention*, 35(4), 537-547. doi:[10.1016/S0001-4575\(02\)00031-3](https://doi.org/10.1016/S0001-4575(02)00031-3)
- Anastasopoulos, P. Ch., Tarko, A. P. & Mannering, F. L. (2008). Tobit analysis of vehicle accident rates on interstate highways. *Accident Analysis & Prevention*, 40(2), 768-775. doi:[10.1016/j.aap.2007.09.006](https://doi.org/10.1016/j.aap.2007.09.006)
- Anselin, L. (1988). *Spatial econometrics: methods and models*. Dordrecht, The Netherlands: Kluwer Academic Publishers.

- Anselin, L. (1990). Some robust approaches to testing and estimation in spatial econometrics. *Regional Science and Urban Economics*, 20(2), 141-163. doi:10.1016/0166-0462(90)90001-J
- Anselin, L. (2005). *Exploring spatial data with GeoDa: A workbook*. 165-212. Center for Spatially Integrated Social Science. University of Illinois, Urbana-Champaign. Retrieved January 20, 2013 from <http://www.csiss.org/> or <http://sal.agecon.uiuc.edu/>
- Anselin, L. & Griffith, D. A. (1988). Do spatial effects really matter in regression analysis? *Papers of the Regional Science Association*, 65(1), 19-33. doi:10.1111/j.1435-5597.1988.tb01155.x
- Dorling, D. (1996). *Area cartograms: Their use and creation*. Concepts and Techniques in Modern Geography series no. 59. Norwich: University of East Anglia.
- Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research*, 40(5), 341-351. doi:10.1016/j.jsr.2009.07.006
- Fridstrøm, L. & Ingebrigtsen, S. (1991). An aggregate accident model based on pooled regional time series data. *Accident Analysis & Prevention*, 23(5), 363-378. doi:10.1016/0001-4575(91)90057-C
- Goh, K. C. K., Currie, G., Sarvi, M. & Logan, D. (2014). Bus accident analysis of routes with/without bus priority. *Accident Analysis & Prevention*, 65, 18-27. doi:10.1016/j.aap.2013.12.002
- Griffith, D. A. (2009). Spatially autoregressive models. In R. Kitchin & N. Thrift (Eds.) *International Encyclopedia of Human Geography* (pp. 396-402). Oxford: Elsevier.
- Guo, F., Wang, X. & Abdel-Aty, M. A. (2010). Modelling signalized intersection safety with corridor-level spatial correlations. *Accident Analysis & Prevention*, 42(1), 84-92. doi:10.1016/j.aap.2009.07.005
- Jegede, F. J. (1988). Spatio-temporal analysis of road traffic accidents in Oyo State, Nigeria. *Accident Analysis & Prevention*, 20(3), 227-243. doi:10.1016/0001-4575(88)90007-3
- Jiang, X., Abdel-Aty, M. & Alamili, S. (2014). Application of Poisson random effect models for highway network screening. *Accident Analysis & Prevention*, 63, 74-82. doi:10.1016/j.aap.2013.10.029

SPATIAL ANALYTICAL FRAMEWORK

- Jones, B., Janssen, L. & Mannering, F. (1991). Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis & Prevention*, 23(4), 239-255. doi:[10.1016/0001-4575\(91\)90003-N](https://doi.org/10.1016/0001-4575(91)90003-N)
- Jovanis, P. & Chang, H. (1986). Modelling the relationship of accidents to miles traveled. *Transportation Research Record*, 1068, 42-51.
- Kim, K., Brunner, I. & Yamashita, E. (2006). Influence of land use, population, employment, and economic activity on accidents. *Transportation Research Record: Journal of Transport Research Board*, 1953, 56-64. doi:[10.3141/1953-07](https://doi.org/10.3141/1953-07)
- Korter, G. O., Olubusoye, O. E. & Salisu, A. A. (2014). Spatial modelling of road traffic accidents in Oyo State, Nigeria. *African Journal of Sustainable Development (AJSD)*, 4 (2), 99-118.
- LeSage, J. & Pace, R. K. (2009). *Introduction to spatial econometrics*. Boca Raton: Chapman and Hall/CRC.
- Levine, N., Kim, K. E. & Nitz, L. H. (1995a). Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns. *Accident Analysis & Prevention*, 27(5), 663-674. doi:[10.1016/0001-4575\(95\)00017-T](https://doi.org/10.1016/0001-4575(95)00017-T)
- Levine, N., Kim, K. E. & Nitz, L. H. (1995b). Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators. *Accident Analysis & Prevention*, 27(5), 675-685. doi:[10.1016/0001-4575\(95\)00018-U](https://doi.org/10.1016/0001-4575(95)00018-U)
- Lord, D. & Miranda-Moreno, L. F. (2008). Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson–gamma models for modeling motor vehicle crashes: A Bayesian perspective. *Safety Science*, 46(5), 751-770. doi:[10.1016/j.ssci.2007.03.005](https://doi.org/10.1016/j.ssci.2007.03.005)
- Maher, M. J. & Mountain, L. J. (1988). The identification of accident blackspots: A comparison of current methods. *Accident Analysis & Prevention*, 20(2), 143-151. doi:[10.1016/0001-4575\(88\)90031-0](https://doi.org/10.1016/0001-4575(88)90031-0)
- McGill, R., Tukey, J. W. & Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, 32(1), 12-16. doi:[10.1080/00031305.1978.10479236](https://doi.org/10.1080/00031305.1978.10479236)
- Miaou, S.-P. & Lord, D. (2003). Modelling traffic crash-flow relationship for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes. *Transportation Research Record: Journal of Transport Research Board* 1840, 31-40. doi:[10.3141/1840-04](https://doi.org/10.3141/1840-04)
- Milton, J. & Mannering, F. (1998). The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies. *Transportation*, 25(4), 395-413. doi:[10.1023/A:1005095725001](https://doi.org/10.1023/A:1005095725001)

- Noland, R.B. & Quddus, M.A. (2005). Congestion and safety: A spatial analysis of London. *Transportation Research Part A: Policy and Practice*, 39(7/9), 737-754. doi:[10.1016/j.tra.2005.02.022](https://doi.org/10.1016/j.tra.2005.02.022)
- Okamoto, H. & Koshi, M. (1989). A method to cope with the random errors of observed accident rates in regression analysis. *Accident Analysis & Prevention*, 21(4), 317-332. doi:[10.1016/0001-4575\(89\)90023-7](https://doi.org/10.1016/0001-4575(89)90023-7)
- Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data. *Accident Analysis & Prevention*, 40(4), 1486-1497. doi:[10.1016/j.aap.2008.03.009](https://doi.org/10.1016/j.aap.2008.03.009)
- Roque, C. & Cardoso, J. L. (2014). Investigating the relationship between run-off the-road crash frequency and traffic flow through different functional forms. *Accident Analysis and Prevention*, 63, 121-132. doi:[10.1016/j.aap.2013.10.034](https://doi.org/10.1016/j.aap.2013.10.034)
- Shanker, V., Mannering, F. & Barfield, W. (1995). Effect of roadway geometrics and environmental factors on rural accident frequencies. *Accident Analysis & Prevention*, 27(3), 371-389. doi:[10.1016/0001-4575\(94\)00078-Z](https://doi.org/10.1016/0001-4575(94)00078-Z)
- Shefer, D. & Rietveld, P. (1997). Congestion and safety on highways: Towards an analytical model. *Urban Studies*, 34 (4), 679-692. doi:[10.1080/0042098975970](https://doi.org/10.1080/0042098975970)
- Wang, C., Quddus, M. A. & Ison, S. G. (2009). Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England. *Accident Analysis & Prevention*, 41(4), 798-808. doi:[10.1016/j.aap.2009.04.002](https://doi.org/10.1016/j.aap.2009.04.002)
- Ye, X., Pendyala, R. M., Washington, S. P., Konduri, K., & Oh, J. (2009). A simultaneous equations model of crash frequency by collision type for rural intersections. *Safety Science*, 47(3), 443-452. doi:[10.1016/j.ssci.2008.06.007](https://doi.org/10.1016/j.ssci.2008.06.007)
- Yu, R. & Abdel-Aty, M. (2013). Investigating different approaches to develop information priors in hierarchical Bayesian safety performance functions. *Accident Analysis & Prevention*, 56, 51-58. doi:[10.1016/j.aap.2013.03.023](https://doi.org/10.1016/j.aap.2013.03.023)
- Yu, R., Abdel-Aty, M., & Ahmed, M. (2013). Bayesian random effect models incorporating real-time weather and traffic data to investigate mountainous freeway hazardous factors. *Accident Analysis & Prevention*, 50, 371-376. doi:[10.1016/j.aap.2012.05.011](https://doi.org/10.1016/j.aap.2012.05.011)

Generalized Singular Value Decomposition With Additive Components

Stan Lipovetsky
GfK North America
Minneapolis, MN

The singular value decomposition (SVD) technique is extended to incorporate the additive components for approximation of a rectangular matrix by the outer products of vectors. While dual vectors of the regular SVD can be expressed one via linear transformation of the other, the modified SVD corresponds to the general linear transformation with the additive part. The method obtained can be related to the family of principal component and correspondence analyses, and can be reduced to an eigenproblem of a specific transformation of a data matrix. This technique is applied to constructing dual eigenvectors for data visualizing in a two dimensional space.

Keywords: singular value decomposition, general linear transformation, principal components, dual eigenvectors, perceptual mapping

---

Introduction

Data visualization, also known as perceptual mapping, is a representation of multivariate observations in two dimensional plots. It can be performed with various techniques, particularly principal component analysis (PCA) and factor analysis (Dillon and Goldstein, 1984; Seber, 1984; Falissard, 1999), multidimensional scaling (Chambers et al., 1983; Cleveland, 1985), and singular value decomposition (SVD) techniques (Golub and van Loan, 1983). These methods are applied for bi-plotting when both observations and variables are represented in the same graph (Gabriel 1971; Gabriel and Odoroff, 1990). Perceptual mapping includes methods of correspondence analysis or dual scaling for categorical data (Nishisato, 1980; Greenacre, 1984; Carroll et al., 1986; Benzecri, 1992), canonical correlation analysis and MANOVA (Novak, 1995; Tishler and Lipovetsky, 1996), and many-way eigenvector analysis (Cooper, 1988; Carlier and Kroonenberg, 1996; Lipovetsky, 1994). Various additional techniques are also used for specific problems in PCA, SVD, and perceptual mapping (Kaciak

---

Dr. Lipovetsky is a Senior Research Director in the Marketing Sciences. Email him at: stan.lipovetsky@gfk.com.

and Louviere, 1990; Sinha and DeSarbo, 1998; Steenkamp et al., 1994; Shugan, 1987; Lipovetsky and Conklin, 2001, 2003, 2004, 2005; Lipovetsky, 2009).

Projection of multivariate observation in a two dimensional space is usually performed by PCA or SVD techniques. In a PCA approach, the variables are centered by their means (and often normalized by their standard deviations), and in an SVD approach the original matrices of data are processed. Both PCA and SVD solve some eigenproblems and, from the several main eigenvectors (those correspond to maximum eigenvalues), a couple of vectors are picked for use as plot axes. In PCA, those are usually the 1st and 2nd or 2nd and 3rd main eigenvectors. In SVD, especially when it is used for correspondence analysis, the 1st eigenvector represents the mean values or their proportions among the variables, so the 2nd and the 3rd eigenvectors are applied.

A general linear transformation, including additive part, to construct the dual vectors for approximation of a data matrix. It can be called the Generalized SVD, or GSVD. This technique can be seen as another member of the family of SVD and PCA methods. GSVD corresponds to the general linear transformation of the axes of a plot, including shift of the beginning of coordinates and change of scales' units.

Despite problems in interpretation of the perceptual maps, the visualization of data is needed and widely used in applied statistical analysis. As it is noted in Jones and Pearce (2000), "In America visually-oriented, quantitatively illiterate culture, images have a great deal of power, so if a picture is today worth a thousand words, it must be worth at least a billion numbers". It could be particularly true for multivariate data and its statistical consideration (see also Lipovetsky and Mandel, 2009).

SVD and PCA

Consider a matrix approximation by an outer product of two dual vectors. Let $\mathbf{x}$ denote a data matrix of m by n order, with elements x_{ij} of i th observations $i = (1, \dots, m)$ by j th variables $j = (1, \dots, n)$. A model of matrix approximation by two vectors is

$$x_{ij} = \lambda b_i a_j + \varepsilon_{ij} , \quad (1)$$

where b_i and a_j are elements of vectors $\mathbf{b}$ and $\mathbf{a}$ of m th and n th order, respectively, λ is a normalizing parameter, and $[\varepsilon_{ij}]$ is a matrix of residuals. This equation in matrix form is

$$\mathbf{x} = \lambda \mathbf{b} \mathbf{a}' + \boldsymbol{\varepsilon} , \quad (2)$$

GENERALIZED SINGULAR VALUE DECOMPOSITION

with $\mathbf{b}$ as a column-vector and $\mathbf{a}'$ as a row-vector (prime denotes transposition). The least squares (LS) procedure for determining the vectors $\mathbf{a}$ and $\mathbf{b}$ by minimizing deviations $\mathbf{\varepsilon}$, also known as Eckart-Young approximation (Eckart and Young, 1936; Lipovetsky and Tishler, 1994), is defined by the objective:

$$S = \|\mathbf{\varepsilon}\|^2 = \|\mathbf{x} - \lambda \mathbf{b} \mathbf{a}\|^2 = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \lambda b_i a_j)^2 \rightarrow \min . \quad (3)$$

From the first-order conditions

$$\partial S / \partial b_i = 0, \quad \partial S / \partial a_j = 0 , \quad (4)$$

a system of equations yields

$$\begin{cases} \sum_{j=1}^n x_{ij} a_j = \lambda b_i \sum_{j=1}^n a_j^2 & , \quad i = 1, \dots, m , \\ \sum_{i=1}^m x_{ij} b_i = \lambda a_j \sum_{i=1}^m b_i^2 & , \quad j = 1, \dots, n . \end{cases} \quad (5)$$

Normalizing vectors by their scalar norm equal one,

$$\sum_{j=1}^n a_j^2 = 1, \quad \sum_{i=1}^m b_i^2 = 1 , \quad (6)$$

we represent (5) in a matrix form as follows:

$$\mathbf{x} \mathbf{a} = \lambda \mathbf{b}, \quad \mathbf{x}' \mathbf{b} = \lambda \mathbf{a} . \quad (7)$$

Substituting one of the equations in (7) into the other, we get two problems:

$$(\mathbf{x}' \mathbf{x}) \mathbf{a} = \lambda^2 \mathbf{a} , \quad (8)$$

that is the eigenproblem for variables aggregate, and

$$(\mathbf{x} \mathbf{x}') \mathbf{b} = \lambda^2 \mathbf{b} , \quad (9)$$

that is the dual eigenproblem for the observations aggregate. First eigenvectors $\mathbf{a}$ and $\mathbf{b}$ for maximum eigenvalue λ^2 in (7)-(8) solve the problem (1)-(2) for approximation of a matrix by dual vectors' outer product.

If $\mathbf{x}$ is a matrix of the standardized (centered and normalized) variables, then $\mathbf{x}'\mathbf{x}$ coincides with the correlation matrix, and the problem (8) coincides with PCA. Indeed, in PCA we construct a linear combination of variables

$$\mathbf{y} = \mathbf{x}\mathbf{a} , \quad (10)$$

and maximize its variance

$$\text{var}(\mathbf{y}) = \mathbf{y}'\mathbf{y} = \mathbf{a}'\mathbf{x}'\mathbf{x}\mathbf{a} \rightarrow \max . \quad (11)$$

With normalizing restriction (6) for vector $\mathbf{a}$, we represent (11) as a conditional objective

$$\mathbf{a}(\mathbf{x}'\mathbf{x})\mathbf{a} + \mu(\mathbf{a}'\mathbf{a} - 1) \rightarrow \max , \quad (12)$$

with a Lagrange term μ . Maximizing (11) due to (4), we obtain the same SVD eigenproblem (8) with $\mu = \lambda^2$. Comparison of (7) and (10) shows that $\mathbf{y} = \lambda\mathbf{b}$, so the aggregator y of PCA is equal to the dual SVD vector $\mathbf{b}$ normalized by λ^2 . Similarly, in a dual PCA approach, we can consider a linear combination of observations

$$\mathbf{z} = \mathbf{x}'\mathbf{b} , \quad (13)$$

and maximize its variance; that yields the SVD eigenproblem (9) and the relation $\mathbf{z} = \lambda\mathbf{a}$ between PCA aggregator $\mathbf{z}$ and SVD dual vector $\mathbf{a}$.

The SVD (7) and PCA relations (10) and (13) show that every element of an eigenvector is a linear combination of the elements of the dual eigenvector. In some analogies, it corresponds to a linear model without intercept. Let us consider now a model with intercept, meaning more general linear relations between dual eigenvectors.

Generalized SVD

In place of the model of matrix approximation by outer product of two vectors (1), let us take a general model with linear and mixed effects:

GENERALIZED SINGULAR VALUE DECOMPOSITION

$$x_{ij} = \mu + b_i + a_j + \lambda b_i a_j + \varepsilon_{ij} . \quad (14)$$

As it is known in the fitting of two-way tables in the analysis of variance (Tukey, 1977; Emerson & Wong, 1985), such a model can be re-arranged to the multiplicative fit model with the additive constant

$$x_{ij} = \nu + \lambda \beta_i \alpha_j + \varepsilon_{ij} , \quad (15)$$

where the relations between the parameters used in (14) and in (15) are as follows:

$$\nu = \mu - \lambda^{-1}, \quad \beta_i = b_i + \lambda^{-1}, \quad \alpha_j = a_j + \lambda^{-1} . \quad (16)$$

The model (15) is the extension of the regular SVD model (1) by the constant item ν . If the values of parameters ν , λ , $\boldsymbol{\beta}$, and $\boldsymbol{\alpha}$ in the model (15) are estimated, then by (16) we obtain all the parameters μ , λ , $\mathbf{b}$, and $\mathbf{a}$ of the model (14) as well.

Take the model (15) and use it in the objective (3); that becomes

$$S = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{x} - \nu - \lambda \boldsymbol{\beta} \boldsymbol{\alpha}'\|^2 = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \nu - \lambda \beta_i \alpha_j)^2 \rightarrow \min . \quad (17)$$

From the condition $\partial S / \partial \nu = 0$, we get the expression for the constant

$$\nu = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n x_{ij} - \lambda \left(\frac{1}{m} \sum_{i=1}^m \beta_i \right) \left(\frac{1}{n} \sum_{j=1}^n \alpha_j \right) \equiv \bar{\mathbf{x}} - \lambda \bar{\boldsymbol{\beta}} \bar{\boldsymbol{\alpha}} , \quad (18)$$

so it equals the mean value by all the data in matrix $\mathbf{x}$ minus the interaction part of the mean values of the $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ dual vectors. Substituting (18) into (15), we represent this model as

$$\tilde{x}_{ij} = \lambda \beta_i \alpha_j - \lambda \bar{\boldsymbol{\beta}} \bar{\boldsymbol{\alpha}} + \varepsilon_{ij} \quad (19)$$

with the matrix of deviation from the total mean defined as

$$\tilde{x}_{ij} = x_{ij} - \bar{\mathbf{x}} . \quad (20)$$

To reduce the model (19) to the regular SVD model (1), we impose the constraint that both mean values of β and α equal zero:

$$\sum_{j=1}^n \alpha_j = 0, \quad \sum_{i=1}^m \beta_i = 0. \quad (21)$$

Such restrictions are used in the analysis of variance with non-additive fits (Mandel, 1969, 1971) and in factor analysis (Gollob, 1968).

With help of the relations (18)-(21), we represent (17) as a conditional objective

$$\begin{aligned} S = \|\mathbf{e}\|^2 &= \|\tilde{\mathbf{x}} - \lambda \beta \alpha'\|^2 + 2\rho(\beta' \mathbf{e}_m) + 2\tau(\alpha' \mathbf{e}_n) \\ &= \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \lambda \beta_i \alpha_j)^2 + 2\rho \sum_{i=1}^m \beta_i + 2\tau \sum_{j=1}^n \alpha_j \rightarrow \min \end{aligned} \quad (22)$$

where ρ and τ are Lagrange multipliers for incorporating conditions (21) into the objective (17), and $\mathbf{e}_m$ and $\mathbf{e}_n$ denote the uniform vector-columns of order m and n , respectively. From the first order conditions (4) applied to the vectors β and α in (22), a system of equations yields:

$$\begin{cases} \sum_{j=1}^n \tilde{x}_{ij} \alpha_j = \lambda \beta_i \sum_{j=1}^n \alpha_j^2 + \rho / \lambda, & i = 1, \dots, m, \\ \sum_{i=1}^m \tilde{x}_{ij} \beta_i = \lambda \alpha_j \sum_{i=1}^m \beta_i^2 + \tau / \lambda, & j = 1, \dots, n. \end{cases} \quad (23)$$

Comparison with (5) shows that equation (23) corresponds to the general linear transformation of one vector to another with additive constants.

Multiplying the first group of equations (23) by β_i and summing over the index i , or the second group of equations (23) by α_j and summing over j , and taking into account conditions (21), we get expression for the term λ :

$$\lambda = \frac{\sum_i \sum_j \tilde{x}_{ij} \beta_i \alpha_j}{\sum_i \beta_i^2 \sum_j \alpha_j^2}. \quad (24)$$

GENERALIZED SINGULAR VALUE DECOMPOSITION

Summing the first group of equations (23) by index i , the second group of equations (23) by index j , and using conditions (21), we obtain expressions for the terms ρ and τ :

$$\rho = \lambda \sum_{j=1}^n \left(\frac{1}{m} \sum_{i=1}^m \tilde{x}_{ij} \right) \alpha_j, \quad \tau = \lambda \sum_{i=1}^m \left(\frac{1}{n} \sum_{j=1}^n \tilde{x}_{ij} \right) \beta_i. \quad (25)$$

Substituting (25) into (23) yields the following system of equations:

$$\begin{cases} \sum_{j=1}^n \left(\tilde{x}_{ij} - \frac{1}{m} \sum_{i=1}^m \tilde{x}_{ij} \right) \alpha_j = \lambda \beta_i \sum_{j=1}^n \alpha_j^2, & i=1, \dots, m, \\ \sum_{i=1}^m \left(\tilde{x}_{ij} - \frac{1}{n} \sum_{j=1}^n \tilde{x}_{ij} \right) \beta_j = \lambda \alpha_j \sum_{i=1}^m \beta_i^2, & j=1, \dots, n. \end{cases} \quad (26)$$

Using definition (20) we can re-write expressions in the parentheses in (26) via the original data matrix:

$$\begin{cases} \tilde{x}_{ij} - \frac{1}{m} \sum_{i=1}^m \tilde{x}_{ij} = x_{ij} - \bar{x} - \frac{1}{m} \sum_{i=1}^m (x_{ij} - \bar{x}) = x_{ij} - c_j, \\ \tilde{x}_{ij} - \frac{1}{n} \sum_{j=1}^n \tilde{x}_{ij} = x_{ij} - \bar{x} - \frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{x}) = x_{ij} - r_i, \end{cases} \quad (27)$$

where c_j and r_i denote, respectively, the column mean (average by rows) and row mean (average by columns) values of the matrix $[x_{ij}]$:

$$c_j = \frac{1}{m} \sum_{i=1}^m x_{ij}, \quad r_i = \frac{1}{n} \sum_{j=1}^n x_{ij}. \quad (28)$$

In a matrix $\mathbf{x}$, rows often correspond to observations and columns to attributes. Then in (27), the first matrix $[x_{ij} - c_j]$ consists of observations centered and the second matrix $[x_{ij} - r_i]$ consists of attributes centered. Denote these two matrices as $\mathbf{Y}$ and $\mathbf{Z}$, so in matrix form the expressions in (27) are

$$\mathbf{Y} = \mathbf{x} - \mathbf{e}_m \mathbf{c}', \quad \mathbf{Z} = \mathbf{x} - \mathbf{r} \mathbf{e}_n', \quad (29)$$

where uniform vectors $\mathbf{e}$ are as in (22), $\mathbf{c}$ and $\mathbf{r}$ are vector-columns of the order n and m , respectively, with the elements of (28), and the prime denoting transposition.

Using (27)-(29) we represent the system (26) in matrix notation:

$$\mathbf{Y}\boldsymbol{\alpha} = \lambda(\boldsymbol{\alpha}'\boldsymbol{\alpha})\boldsymbol{\beta} \quad , \quad \mathbf{Z}'\boldsymbol{\beta} = \lambda(\boldsymbol{\beta}'\boldsymbol{\beta})\boldsymbol{\alpha} \quad . \quad (30)$$

Normalizing vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by scalar norm (6), we simplify (30) to the system

$$\mathbf{Y}\boldsymbol{\alpha} = \lambda\boldsymbol{\beta} \quad , \quad \mathbf{Z}'\boldsymbol{\beta} = \lambda\boldsymbol{\alpha} \quad . \quad (31)$$

Substituting one of the equations in (31) into the other yields the problems

$$(\mathbf{Z}'\mathbf{Y})\boldsymbol{\alpha} = \lambda^2\boldsymbol{\alpha} \quad , \quad (\mathbf{Y}\mathbf{Z}')\boldsymbol{\beta} = \lambda^2\boldsymbol{\beta} \quad . \quad (32)$$

Relations (31) and (32) define the generalized SVD method. The eigenproblems (32) yield the attributes' and observations' aggregates corresponded to generalizations of the regular SVD solutions (8)-(9). First eigenvectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ for maximum eigenvalue λ^2 in (32) solve the problem (17) or (22) for approximation of a matrix by outer product of vectors with additive component.

The GSVD problems (32) can be represented, using (29), more explicitly:

$$\begin{aligned} ((\mathbf{x}' - \mathbf{e}_n \mathbf{r}')(\mathbf{x} - \mathbf{e}_m \mathbf{c}'))\boldsymbol{\alpha} &= \lambda^2\boldsymbol{\alpha} \quad , \\ ((\mathbf{x} - \mathbf{e}_m \mathbf{c}')(\mathbf{x}' - \mathbf{e}_n \mathbf{r}'))\boldsymbol{\beta} &= \lambda^2\boldsymbol{\beta} \quad . \end{aligned} \quad (33)$$

It then becomes clear that problems in (33) belong to the family of SVD and PCA techniques while differing from both of them. The regular SVD (7)-(9) operates with a matrix $\mathbf{x}$ and the same matrix transposed $\mathbf{x}'$, while in the GSVD (31)-(33) there actually are two matrices corresponded to data centering in different directions (29). Suppose matrix $\mathbf{x}$ consists of centered and normalized variables and $\mathbf{x}'\mathbf{x}$ coincides with the correlation matrix. In this case, due to (29), $\mathbf{Y} = \mathbf{x}$ but $\mathbf{Z} \neq \mathbf{Y}$, and the problems (32)-(33) cannot be reduced to PCA.

For numerical evaluations it could be convenient to express vector-columns $\mathbf{c}$ and $\mathbf{r}$ (28) in matrix notation:

$$\mathbf{c} = \frac{1}{m} \mathbf{x}' \mathbf{e}_m \quad , \quad \mathbf{r} = \frac{1}{n} \mathbf{x} \mathbf{e}_n \quad . \quad (34)$$

GENERALIZED SINGULAR VALUE DECOMPOSITION

Using (34), we represent the problems in (33) directly via the data matrix:

$$\begin{aligned} \left(\left(\mathbf{I}_n - \frac{1}{n} \mathbf{e}_n \mathbf{e}_n' \right) \mathbf{x}' \left(\mathbf{I}_m - \frac{1}{m} \mathbf{e}_m \mathbf{e}_m' \right) \mathbf{x} \right) \boldsymbol{\alpha} &= \lambda^2 \boldsymbol{\alpha}, \\ \left(\left(\mathbf{I}_m - \frac{1}{m} \mathbf{e}_m \mathbf{e}_m' \right) \mathbf{x} \left(\mathbf{I}_n - \frac{1}{n} \mathbf{e}_n \mathbf{e}_n' \right) \mathbf{x}' \right) \boldsymbol{\beta} &= \lambda^2 \boldsymbol{\beta}, \end{aligned} \quad (35)$$

where $\mathbf{I}_n$ and $\mathbf{I}_m$ are identity matrices of the order n and m , respectively. We see that numerical estimations for GSVD problems (35) are not much more difficult than for the regular SVD (8)-(9). It is sufficient to solve just one eigenproblem of the smaller dimension – usually the number of variables n is less than the number of observations m , so we can solve the first problem of (35). The dual solution can then be found by the linear transformation (31).

Although an eigenvector in a separated eigenproblem is defined up to an arbitrary sign, the vectors of dual solutions have directions defined by their dual transformation (formula (7) for SVD, (10) and (13) for PCA, or (31) for GSVD). It means that sign could be changed to the opposite simultaneously in each pair of dual vectors.

Numerical Examples

Consider a matrix of proportions from an actual marketing research project for six brands, $X1, X2, \dots, X6$ and nine attributes, $A, B, \dots, I$, presented in Table 1.

Table 1. Example I: Matrix of proportions.

	X1	X2	X3	X4	X5	X6
A	0.249	0.166	0.317	0.119	0.059	0.077
B	0.266	0.283	0.345	0.293	0.154	0.228
C	0.217	0.203	0.279	0.249	0.148	0.214
D	0.127	0.051	0.129	0.231	0.054	0.206
E	0.211	0.222	0.335	0.211	0.153	0.235
F	0.278	0.246	0.31	0.285	0.181	0.26
G	0.266	0.289	0.438	0.378	0.257	0.426
H	0.259	0.263	0.349	0.257	0.164	0.265
I	0.246	0.302	0.444	0.167	0.222	0.232

Table 2. Example I: Pairs of dual vectors for SVD, PCA, and GSVD.

	SVD-vectors		PCA-vectors		GSVD-vectors	
	$\mathbf{a2}$	$\mathbf{a3}$	$\mathbf{a2}$	$\mathbf{a3}$	$\boldsymbol{\alpha1}$	$\boldsymbol{\alpha2}$
<i>X1</i>	0.22	-0.61	0.22	-0.59	0.21	-0.70
<i>X2</i>	0.30	0.16	0.31	-0.30	0.34	-0.01
<i>X3</i>	0.47	0.10	0.48	0.26	0.51	0.36
<i>X4</i>	-0.58	-0.42	-0.58	-0.53	-0.57	-0.34
<i>X5</i>	-0.02	0.53	-0.01	0.27	0.01	0.22
<i>X6</i>	-0.55	0.35	-0.54	0.38	-0.50	0.47
	$\mathbf{b2}$	$\mathbf{b3}$	$\mathbf{b2}$	$\mathbf{b3}$	$\boldsymbol{\beta1}$	$\boldsymbol{\beta2}$
<i>A</i>	0.54	-0.56	0.53	-0.14	0.48	-0.46
<i>B</i>	0.04	-0.27	0.04	-0.56	0.04	-0.25
<i>C</i>	-0.09	-0.13	-0.09	-0.08	-0.10	-0.13
<i>D</i>	-0.54	-0.35	-0.56	0.28	-0.61	-0.19
<i>E</i>	0.07	0.12	0.07	0.30	0.07	0.20
<i>F</i>	-0.11	-0.19	-0.11	-0.39	-0.10	-0.24
<i>G</i>	-0.39	0.40	-0.38	0.19	-0.32	0.57
<i>H</i>	0.01	-0.04	0.02	-0.13	0.03	0.01
<i>I</i>	0.48	0.51	0.49	0.53	0.51	0.49

The results obtained in SVD, PCA, and GSVD solutions are presented in Table 2, where all of the vectors are normalized by the totals of the squared elements equal one. For the SVD (7)-(9) and PCA (10)-(13) solutions, we take the second ($\mathbf{a2}$, $\mathbf{b2}$) and the third ($\mathbf{a3}$, $\mathbf{b3}$) pairs of dual vectors (because the first dimension of means is related to intensity rather than to structure). For the generalized solution (35) we take its first ($\boldsymbol{\alpha1}$, $\boldsymbol{\beta1}$) and second ($\boldsymbol{\alpha2}$, $\boldsymbol{\beta2}$) pairs of dual vectors (because due to the construction of this method we have already reduced the data to just the structure among the items in both directions).

Comparison among the vectors $\mathbf{a2}$ for SVD, $\mathbf{a2}$ for PCA, and $\boldsymbol{\alpha1}$ for GSVD shows that they are very close to one another. Their dual vectors are also similar among themselves, although $\boldsymbol{\beta1}$ for GSVD solutions differs from the two other very close solutions $\mathbf{b2}$ for SVD and $\mathbf{b2}$ for PCA. The next solutions, $\mathbf{a3}$ for SVD, $\mathbf{a3}$ for PCA, and $\boldsymbol{\beta2}$ for GSVD, are already not so similar among themselves; they can differ by value and by sign, too. The same concerns their dual vectors $\mathbf{b3}$ for SVD,

GENERALIZED SINGULAR VALUE DECOMPOSITION

b3 for PCA, and **$\beta 2$** for GSVD. However, the biggest coefficients are always of the same sign within each group of the vectors compared.

Another example presents real data for ten brands and fifteen attributes (those are: good value, natural, variety of flavors, I would pay more, good to serve to guests, reenergizes, innovative, brand I trust, cares for environment, socially responsible, served in selected restaurants, authentic, symbol of conviviality, high quality, my brand). The proportions by choices are given in Table 3.

The results of SVD, PCA, and GSVD solutions are presented in Table 4 and are constructed similarly to those described above in Table 2, with the second (**a2**, **b2**) and third (**a3**, **b3**) pairs of dual vectors for SVD and PCA and the first (**$\alpha 1$** , **$\beta 1$**) and second (**$\alpha 2$** , **$\beta 2$**) pairs of dual vectors for GSVD. Again, the vectors **a2** for SVD and PCA and **$\alpha 1$** for GSVD are very similar. The corresponding dual vectors are less similar among themselves. The next solutions, **a3** for SVD and PCA and **$\alpha 2$** for GSVD, differ more; the same concerns their dual vectors, **b3** for SVD and PCA and **$\beta 2$** for GSVD. The reason for the differences is easy to explain by recalling that the total of the elements in any SVD vector does not equal zero while **b**-vectors in PCA have zero totals because of the centering of data in columns. In

Table 3. Example II: Matrix of proportions.

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
A	0.166	0.194	0.198	0.085	0.108	0.036	0.099	0.086	0.119	0.093
B	0.250	0.282	0.259	0.071	0.085	0.068	0.102	0.085	0.083	0.068
C	0.143	0.168	0.281	0.078	0.131	0.040	0.072	0.108	0.077	0.075
D	0.196	0.214	0.241	0.052	0.063	0.037	0.049	0.044	0.045	0.044
E	0.281	0.332	0.319	0.082	0.110	0.055	0.097	0.099	0.078	0.063
F	0.213	0.260	0.265	0.090	0.110	0.056	0.086	0.088	0.081	0.065
G	0.164	0.213	0.239	0.069	0.094	0.058	0.067	0.086	0.047	0.051
H	0.285	0.335	0.326	0.094	0.123	0.055	0.111	0.112	0.095	0.078
I	0.145	0.191	0.209	0.058	0.071	0.048	0.081	0.079	0.060	0.048
J	0.166	0.222	0.218	0.056	0.077	0.049	0.073	0.072	0.059	0.045
K	0.256	0.304	0.260	0.049	0.046	0.033	0.056	0.058	0.040	0.037
L	0.270	0.166	0.163	0.045	0.047	0.035	0.033	0.032	0.026	0.029
M	0.171	0.187	0.193	0.055	0.063	0.043	0.050	0.054	0.043	0.041
N	0.282	0.349	0.332	0.091	0.107	0.057	0.088	0.100	0.080	0.067
O	0.170	0.213	0.205	0.058	0.083	0.034	0.061	0.074	0.057	0.058

GSVD, the solution for each vector equals zero due to the data centering in both directions. Which of the solutions is more interpretable and preferable depends on a nature of data in the project. If a decision maker is not satisfied with the results of classical SVD or PCA, it is possible to try the new approach of the GSVD.

Table 4. Example II: Pairs of dual vectors for SVD, PCA, and GSVD

	SVD-vectors		PCA-vectors		GSVD-vectors	
	<i>a2</i>	<i>a3</i>	<i>a2</i>	<i>a3</i>	<i>a1</i>	<i>a2</i>
<i>X1</i>	-0.594	0.590	-0.572	0.523	-0.518	0.667
<i>X2</i>	-0.240	-0.420	-0.157	-0.217	-0.607	-0.252
<i>X3</i>	0.213	-0.414	0.254	-0.514	-0.274	-0.603
<i>X4</i>	0.196	0.174	0.204	0.175	0.170	0.085
<i>X5</i>	0.412	0.191	0.420	0.153	0.241	-0.158
<i>X6</i>	0.051	0.022	0.050	0.004	0.166	0.251
<i>X7</i>	0.238	0.170	0.256	0.268	0.152	0.020
<i>X8</i>	0.326	-0.015	0.342	-0.001	0.188	-0.147
<i>X9</i>	0.323	0.368	0.336	0.459	0.233	0.035
<i>X10</i>	0.259	0.269	0.260	0.281	0.248	0.102
	<i>b2</i>	<i>b3</i>	<i>b2</i>	<i>b3</i>	<i>b1</i>	<i>b2</i>
<i>A</i>	0.380	0.556	0.350	0.611	0.384	0.086
<i>B</i>	-0.062	0.167	-0.041	0.246	-0.129	0.115
<i>C</i>	0.525	-0.044	0.498	-0.204	0.366	-0.408
<i>D</i>	-0.136	-0.164	-0.167	-0.264	-0.011	0.026
<i>E</i>	-0.090	-0.082	-0.030	-0.006	-0.355	-0.148
<i>F</i>	0.129	0.059	0.141	0.082	0.009	-0.088
<i>G</i>	0.152	-0.197	0.127	-0.265	0.162	-0.139
<i>H</i>	0.028	0.066	0.098	0.173	-0.310	-0.172
<i>I</i>	0.168	-0.094	0.125	-0.125	0.263	-0.053
<i>J</i>	0.058	-0.151	0.031	-0.155	0.128	-0.047
<i>K</i>	-0.434	-0.338	-0.418	-0.305	-0.386	0.055
<i>L</i>	-0.521	0.630	-0.589	0.442	-0.035	0.807
<i>M</i>	-0.064	-0.003	-0.118	-0.090	0.165	0.164
<i>N</i>	-0.099	-0.195	-0.028	-0.104	-0.407	-0.218
<i>O</i>	0.056	-0.015	0.022	-0.037	0.157	0.018

Summary

A modified singular value decomposition technique is described as a general linear model with an additive part in the mutual representation of the dual vectors one via another. The dual transformation with additional parameters is more flexible to adjusting one direction of the data matrix to another, so the generalized SVD can better express the structure of the simultaneous positioning of a matrix's two directions in the two-dimensional plot. The GSVD technique is reduced to an eigenproblem of a specific structure of the product of two matrices, each centered in one of the directions of the data (35). This eigenproblem does not produce the eigenvectors of intensity that correspond to the first dual vectors in singular value decomposition or in principal component analysis. The GSVD eigenvectors, beginning from the first one, describe the structure of the relations among the directions in the data matrix. The generalized SVD yields a possibility to obtain an additional structure of items' locations on the perceptual map. Comparing maps of different techniques, a researcher can find the more interpretable perceptual map. The GSVD technique enriches the family of SVD and PCA methods for obtaining dual vectors, and can be useful for theoretical investigation and practical applications in numerous problems that use eigenvector or SVD analysis.

References

- Benzecri, J. P. (1992). Correspondence analysis handbook. New York: Marcel Dekker Inc.
- Carlier, A. & Kroonenberg, P. M. (1996). Decompositions and biplots in three-way correspondence analysis. *Psychometrika*, 61(2), 355-373. doi: [10.1007/BF02294344](https://doi.org/10.1007/BF02294344)
- Carroll, J. D., Green, P. E., & Schaffer C. M. (1986). Interpoint distance comparisons in correspondence analysis. *Journal of Marketing Research*, 23(3), 271-280. doi: [10.2307/3151485](https://doi.org/10.2307/3151485)
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Graphical methods for data analysis. Belmont, CA: Wadsworth.
- Cleveland, W. S. (1985). The elements of graphing data. Monterey, CA: Wadsworth.
- Cooper, L. G. (1988) Competitive maps: The structure underlying asymmetric cross elasticities. *Management Science*, 34(6), 707-723. doi: [10.1287/mnsc.34.6.707](https://doi.org/10.1287/mnsc.34.6.707)

- Dillon, W. R. & Goldstein, M. (1984). Multivariate analysis, Methods and applications. New York: Wiley.
- Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(2), 211-218. doi: [10.1007/BF02288367](https://doi.org/10.1007/BF02288367)
- Emerson, J. D. & Wong G. Y. (1985). Resistant nonadditive fits for two-way tables. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (67-124). New York: John Wiley and Sons.
- Falissard, B. (1999). Focused principal component analysis: Looking at a correlation matrix with a particular interest in a given variable. *Journal of Computational and Graphical Statistics*, 8(4), 906-912. doi: [10.1080/10618600.1999.10474855](https://doi.org/10.1080/10618600.1999.10474855)
- Gabriel, K. R. (1971). The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, 58(3), 453-467. doi: [10.1093/biomet/58.3.453](https://doi.org/10.1093/biomet/58.3.453)
- Gabriel, K. R. & Odoroff, C. L. (1990). Biplots in biomedical research. *Statistics in Medicine*, 9(5), 469-485. doi: [10.1002/sim.4780090502](https://doi.org/10.1002/sim.4780090502)
- Gollob, H. F. (1968). A statistical model which combines features of factor analysis and analysis of variance techniques. *Psychometrika*, 33(1), 73-115. doi: [10.1007/BF02289676](https://doi.org/10.1007/BF02289676)
- Golub, G. H. & van Loan, C. F. (1983). Matrix computations. Baltimore: Johns Hopkins University Press.
- Greenacre, M. J. (1984). Theory and application of correspondence analysis. London: Academic Press.
- Jones, R. & Pearce, J. (2000). A postmodern view of fractions and the reciprocals of Fermat primes. *Mathematics Magazine*, 73(2), 83-97. doi: [10.2307/2691078](https://doi.org/10.2307/2691078)
- Kaciak, E., & Louviere J. (1990). Multiple correspondence analysis of multiple choice experiment data. *Journal of Marketing Research*, 27(4), 455-465. doi: [10.2307/3172630](https://doi.org/10.2307/3172630)
- Lipovetsky, S. (1994). Multimode data analysis for decision making. In G. H. Tzeng et al. (Eds.), *Multiple criteria decision making* (275-282). Springer-Verlag.
- Lipovetsky, S. (2009). PCA and SVD with nonnegative loadings. *Pattern Recognition*, 42(1), 68-76. doi: [10.1016/j.patcog.2008.06.025](https://doi.org/10.1016/j.patcog.2008.06.025)
- Lipovetsky, S. & Conklin, M. (2001). Dual priority-antipriority Thurstone scales as AHP eigenvectors. *Engineering Simulation*, 18(2001), 631-648.

GENERALIZED SINGULAR VALUE DECOMPOSITION

- Lipovetsky, S. & Conklin, M. (2003) Dual- and triple-mode matrix approximation and regression modeling. *Applied Stochastic Models in Business and Industry*, 19(4), 291-301. doi: [10.1002/asmb.503](https://doi.org/10.1002/asmb.503)
- Lipovetsky, S. & Conklin, M. (2004). Nonlinear Thurstone scaling via SVD and Gower plots. *International Journal of Operations and Quantitative Management*, 10(4), 259-273.
- Lipovetsky, S. & Conklin, M. (2005). Singular value decomposition in additive, multiplicative, and logistic forms. *Pattern Recognition*, 38(7), 1099-1110. doi: [10.1016/j.patcog.2005.01.010](https://doi.org/10.1016/j.patcog.2005.01.010)
- Lipovetsky, S. & Mandel, I. (2009). How art helps to understand statistics. *Model Assisted Statistics and Applications*, 4(4), 313-324. doi: [10.3233/MAS-2009-0135](https://doi.org/10.3233/MAS-2009-0135)
- Lipovetsky, S. & Tishler, A. (1994). Linear methods in multimode data analysis for decision making. *Computers and Operations Research*, 21(2), 169-183. doi: [10.1016/0305-0548\(94\)90050-7](https://doi.org/10.1016/0305-0548(94)90050-7)
- Mandel J. (1969). A method for fitting empirical surfaces to physical or chemical data. *Technometrics*, 11(3), 411-429. doi: [10.1080/00401706.1969.10490704](https://doi.org/10.1080/00401706.1969.10490704)
- Mandel J. (1971). A new analysis of variance model for non-additive data. *Technometrics*, 13(1), 1-18. doi: [10.1080/00401706.1971.10488751](https://doi.org/10.1080/00401706.1971.10488751)
- Nishisato, S. (1980). Analysis of categorical data: Dual scaling and its applications. Toronto: Toronto University Press.
- Novak, T.P. (1995). MANOVAMAP: Graphical representation of MANOVA in marketing research. *Journal of Marketing Research*, 32(3), 357-374. doi: [10.2307/3151987](https://doi.org/10.2307/3151987)
- Seber, G. A. F. (1984). Multivariate observations. New York: Wiley.
- Shugan, S.M. (1987). Estimating brand positioning maps using supermarket scanning data. *Journal of Marketing Research*, 24(1), 1-18. doi: [10.2307/3151749](https://doi.org/10.2307/3151749)
- Sinha, I., & DeSarbo W. S. (1998). An integrated approach toward the spatial modeling of perceived customer value. *Journal of Marketing Research*, 35(2), 236-249. doi: [10.2307/3151851](https://doi.org/10.2307/3151851)
- Steenkamp, J. B., van Trijp, H., & Ten Berge, J. (1994). Perceptual mapping based on idiosyncratic sets. *Journal of Marketing Research*, 31(1), 15-27. doi: [10.2307/3151943](https://doi.org/10.2307/3151943)

STAN LIPOVETSKY

Tishler, A. & Lipovetsky, S. (1996). Canonical correlation analysis for three data sets: A unified framework with applications to management. *Computers and Operations Research*, 23(7), 667-679. doi: [10.1016/0305-0548\(95\)00067-4](https://doi.org/10.1016/0305-0548(95)00067-4)

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley.

Almost Unbiased Estimator Using Known Value of Population Parameter(s) in Sample Surveys

Rajesh Singh
Banaras Hindu University
Varanasi, India

S. B. Gupta
SRMS Institute of
Medical Sciences
Uttar Pradesh, India

Sachin Malik
SRM University Delhi
Sonapat, Haryana, India

An almost unbiased estimator using known value of some population parameter(s) is proposed. A class of estimators is defined which includes Singh and Solanki (2012) and Sahai and Ray (1980), Sisodiya and Dwivedi (1981), Singh, Cauhan, Sawan, and Smarandache (2007), Upadhyaya and Singh (1984), Singh and Tailor (2003) estimators. Under simple random sampling without replacement (SRSWOR) scheme the expressions for bias and mean square error (MSE) are derived. Numerical illustrations are given.

Keywords: Auxiliary information, bias, mean square error, unbiased estimator

---

Introduction

The precision of the estimates of the population mean or total of the study variable y can be considering improved by the use of known information on an auxiliary variable x which is highly correlated with the study variable y . Consider a finite population $U = U_1, U_2, \dots, U_N$ of N units. Let y and x stand for the variable under study and auxiliary variable respectively. Let (y_i, x_i) , $i = 1, 2, \dots, n$ denote the values of the units included in a sample s_n of size n drawn by simple random sampling without replacement (SRSWOR). The auxiliary information has been used in improving the precision of the estimate of a parameter (see Sukhatme, Sukhatme, Sukhatme, & Ashok (1984) and the references cited therein). Among many methods, the ratio and product methods of estimation are good illustrations in this context.

Rajesh Singh is an Assistant Professor in the Department of Statistics. Email him at rsinghstat@gmail.com. Dr. Gupta is in the Department of Community Medicine. Email at dr_sbgupta@rediffmail.com. Sachin Malik in the Department of Community Medicine. Email him at sachinkurava999@gmail.com.

In order to have a survey estimate of the population mean $\bar{Y}$ of the study character y , assuming the knowledge of the population mean $\bar{X}$ of the auxiliary character x , the well-known ratio estimator is

$$t_R = \bar{y} \frac{\bar{X}}{\bar{x}} \quad (1)$$

Bahl and Tuteja (1991) suggested an exponential ratio type estimator as

$$t_{\text{exp}} = \bar{y} \exp \left[\frac{\bar{X} - \bar{x}}{\bar{X} + \bar{x}} \right] \quad (2)$$

Several authors have used prior value of certain population parameter(s) to find more precise estimates. Sisodiya and Dwivedi (1981), Sen (1978) and Upadhyaya and Singh (1984) used the known coefficient of variation (CV) of the auxiliary character for estimating population mean of a study character in ratio method of estimation. The use of prior value of coefficient of kurtosis in estimating the population variance of study character y was first made by Singh et al. (1973) Later used by Singh and Kakran (1993) in the estimation of population mean of study character. Singh and Tailor (2003) proposed a modified ratio estimator by using the known value of correlation coefficient. Kadilar and Cingi (2006) and Singh, Pandey, and Hirano (2008) have suggested modified ratio estimators by using different pairs of known value of population parameter(s).

Under SRSWOR, an almost unbiased estimator for estimating $\bar{Y}$ is proposed. To obtain the bias and MSE,

$$\bar{y} = \bar{Y}(1 + e_0), \quad \bar{x} = \bar{X}(1 + e_1),$$

such that $E(e_0) = E(e_1) = 0$.

$$E(e_0^2) = f_1 C_y^2, \quad E(e_1^2) = f_1 C_x^2, \quad E(e_0 e_1) = f_1 \rho C_y C_x$$

where

$$f_1 = \left(\frac{1}{n} - \frac{1}{N} \right), S_y^2 = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{Y})^2, S_x^2 = \frac{1}{(N-1)} \sum_{i=1}^N (x_i - \bar{X})^2$$

$$C_y = \frac{S_y}{Y}, C_x = \frac{S_x}{X}, K_x = \rho_{yx} \left(\frac{C_y}{C_x} \right), \rho_{yx} = \frac{S_{yx}}{(S_y S_x)}, S_{yx} = \frac{1}{(N-1)} \sum_{i=1}^N (y_i - \bar{Y})(x_i - \bar{X})$$

The Proposed Estimator

Consider the following estimator

$$t_1 = \bar{y} \left(\frac{K_1 \bar{X} + K_2 K_3}{K_1 \bar{x} + K_2 K_3} \right)^\alpha \quad (3)$$

The bias and MSE expressions of the estimator t_1 up to the first order of approximation are, respectively, given by

$$B(t_1) = \bar{y} f_1 C_x^2 \left[\frac{\alpha(\alpha+1)V_1^2}{2} - \alpha V_1 K_x \right] \quad (4)$$

$$\text{MSE}(t_1) = \bar{Y}^2 f_1 \left[C_y^2 + C_x^2 (\alpha^2 V_1^2 + 2V_1 \alpha K_x) \right] \quad (5)$$

Following Singh and Solanki (2012), consider the following estimator

$$t_2 = \bar{y} \left\{ 2 - \left(\frac{\bar{x}}{\bar{X}} \right)^\beta \exp \left[\lambda \left(\frac{(K_4 \bar{X} + K_5) - (K_4 \bar{x} + K_5)}{(K_4 \bar{X} + K_5) + (K_4 \bar{x} + K_5)} \right) \right] \right\} \quad (6)$$

The bias and MSE expressions of the estimator t_2 up to the first order of approximation are, respectively, given by

$$B(t_2) = \bar{y} f_1 C_x^2 \left[\frac{\lambda V_2 \beta}{2} - \frac{\beta(\beta-1)}{2} - \frac{\lambda(\lambda+2)V_2^2}{8} - \beta K_x + \frac{\lambda V_2 K_x}{2} \right] \quad (7)$$

$$\text{MSE}(t_2) = \bar{Y}^2 f_1 \left[C_y^2 + C_x^2 \left(\beta^2 + \frac{\lambda^2 V_2^2}{4} - \beta \lambda V_2 \right) - 2K_x C_x^2 \left(\beta - \frac{\lambda V_2}{2} \right) \right] \quad (8)$$

α, λ , and β are suitable chosen constants. Also K_1, K_3, K_4, K_5 are either real numbers or function of known parameters of the auxiliary variable x such as $C_x, \beta_2(x), \rho_{yx}$ and K_x . K_2 is an integer which takes values +1 and -1 for designing the estimators and

$$\left. \begin{aligned} V_1 &= \frac{K_1 \bar{X}}{K_1 \bar{X} + K_2 K_3} \\ V_2 &= \frac{K_4 \bar{X}}{K_4 \bar{X} + K_5} \end{aligned} \right\}$$

The estimators t_1 and t_2 are biased estimators. In some applications bias is disadvantageous. Following these estimators we have proposed almost unbiased estimator of $\bar{Y}$.

Almost Unbiased Estimator

Suppose

$$t_0 = \bar{y}t_1 = \bar{y} \left(\frac{K_1 \bar{X} + K_2 K_3}{K_1 \bar{x} + K_2 K_3} \right)^\alpha \quad t_2 = \left\{ 2 - \left(\frac{\bar{x}}{\bar{X}} \right)^\beta \exp \left[\lambda \left(\frac{(K_4 \bar{X} + K_5) - (K_4 \bar{x} + K_5)}{(K_4 \bar{X} + K_5) + (K_4 \bar{x} + K_5)} \right) \right] \right\}$$

such that $t_0, t_1, t_2 \in W$, where W denotes the set of all possible estimators for estimating the population mean $\bar{Y}$. By definition, the set W is a linear variety if

$$t_p = \sum_{i=0}^3 w_i t_i \in W \quad (9)$$

such that,

$$\sum_{i=0}^3 w_i = 1 \text{ and } w_i \in R \quad (10)$$

ALMOST UNBIASED ESTIMATOR OF POPULATION PARAMETER

where $w_i (i = 0, 1, 2, 3)$ denotes the constants used for reducing the bias in the class of estimators, H denotes the set of those estimators that can be constructed from $t_i (i = 0, 1, 2, 3)$ and R denotes the set of real numbers.

Expressing t_p in terms of e 's,

$$t_p = \bar{Y} \left[1 + e_0 + w_1 \left(\frac{\alpha(\alpha+1)V_1^2 e_1^2}{2} - \alpha V_1 e_1 - \alpha V_1 e_0 e_1 \right) + w_2 \left(-\beta e_1 - \frac{\beta(\beta-1)e_1^2}{2} + \frac{\lambda V_2 e_1}{2} + \frac{\lambda V_2 \beta e_1^2}{2} - \frac{\lambda(\lambda+2)V_2^2 e_1^2}{8} - \beta e_0 e_1 + \frac{\lambda V_2 e_0 e_1}{2} \right) \right] \quad (11)$$

Subtracting $\bar{Y}$ from both sides of equation (11) and then taking expectation of both sides, the bias of the estimator t_p is obtained up to the first order of approximation, as

$$B(t_p) = \bar{Y} f_1 w_1 C_x^2 \left(\frac{\alpha(\alpha+1)V_1^2}{2} - \alpha V_1 K_x \right) + \bar{Y} f_1 w_2 C_x^2 \left(\frac{\lambda V_2 \beta}{2} - \frac{\beta(\beta-1)}{2} - \frac{\lambda(\lambda+2)V_2^2}{8} - \beta K_x + \frac{\lambda V_2 K_x}{2} \right) \quad (12)$$

From (11), we have

$$(t_p - \bar{Y}) = \bar{Y} \left[e_0 - w_1 \alpha V_1 e_1 - w_2 \left(\beta e_1 + \frac{\lambda V_2 e_1}{2} \right) \right] \quad (13)$$

Squaring both sides of (13) and then taking expectation, the MSE of the estimator t_p up to the first order of approximation is obtained, as

$$\text{MSE}(t_p) = \bar{Y}^2 f_1 \left[C_y^2 + C_x^2 (Q^2 - 2QK_x) \right] \quad (14)$$

which is a minimum when

$$Q = K_x \quad (15)$$

where

$$Q = w_1 \alpha V_1 + w_2 \left(\beta - \frac{\lambda V_2}{2} \right) \quad (16)$$

Putting the value of $Q = K_x$ in (14), the optimum value of estimator as t_p (optimum) is obtained. Thus, the minimum MSE of t_p is given by

$$\min .\text{MSE}(t_p) = \bar{Y}^2 f_1 C_y^2 (1 - \rho_{yx}^2) \quad (17)$$

which is same as that of traditional linear regression estimator.

From (10) and (16), there are two equations and three unknowns. It is not possible to find the unique values for w_i 's, $i = 0, 1, 2$. In order to get unique values of w_i 's, impose the linear restriction

$$\sum_{i=0}^2 w_i B(t_i) = 0 \quad (18)$$

where $B(t_i)$ denotes the bias in the i^{th} estimator.

Equations (10), (16) and (18) can be written in the matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & \alpha V_1 & \beta - \frac{\lambda V_2}{2} \\ 0 & B(t_1) & B(t_2) \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 1 \\ k_x \\ 0 \end{bmatrix} \quad (19)$$

Using (19), the unique values of w_i 's, $i = 0, 1, 2$ are

$$\left. \begin{aligned} w_0 &= w_1 = \frac{\alpha V_1 [\alpha V_1 A_2 - A_1 X_1] - X_1 K_x A_1 - X_2 \alpha V_1 [\alpha V_1 A_2 - A_1 X_1] - \alpha V_1 K_x A_1}{\alpha V_1 [\alpha V_1 A_2 - A_1 X_1]} \\ w_2 &= \frac{X_1 K_x A_1}{\alpha V_1 [\alpha V_1 A_2 - A_1 X_1]} + X_2 \\ w_2 &= \frac{K_x A_1}{[\alpha V_1 A_2 - A_1 X_1]} \end{aligned} \right\}$$

ALMOST UNBIASED ESTIMATOR OF POPULATION PARAMETER

where,

$$\left. \begin{aligned} A_1 &= \frac{\alpha(\alpha+1)V_1^2}{2} - \alpha V_1 K_x \\ A_2 &= \frac{\lambda V_2 \beta}{2} - \frac{\beta(\beta-1)}{2} - \frac{\lambda(\lambda+2)V_2^2}{8} - \beta K_x + \frac{\lambda V_2 K_x}{2} \\ X_1 &= A_1 \left[\beta - \frac{\lambda V_2}{2} \right] \\ X_2 &= \frac{K_x}{\alpha V_1} \end{aligned} \right\}$$

Use of these w_i 's, $i = 0, 1, 2$ remove the bias up to terms of order $o(n^{-1})$ at (9).

Empirical Study

For the empirical study, consider the data sets by Kadilar and Cingi (2006) (population 1) and Khoshnevisan, Singh, Chauhan, Sawan, and Smarandache (2007) (population 2).

Data Statistics

Population	N	n	$\bar{Y}$	$\bar{X}$	C_y	C_x	ρ_{yx}	$\beta_2(x)$
Population 1	106	20	2212.59	27421.7	5.22	2.1	0.86	34.57
Population 2	20	8	19.55	18.8	0.355	0.394	-0.92	3.06

Table 1. Values of w_i

w_i	Population 1	Population 2
w_0	2.104965	3.692323
w_1	-6.48599	1.379436
w_2	5.381022	-4.07176

The percent relative efficiencies (PRE) of various estimators of $\bar{Y}$ are computed and presented in Table 2 below.

Table 2. PRE of different estimators of $\bar{Y}$ with respect to $\bar{y}$

Choice of scalars											Estimator	PRE (POPI)	PRE (POPII)
w_0	w_1	w_2	K_1	K_2	K_3	K_4	K_5	α	β	λ			
1	0	0									$\bar{y}$	100	100
0	1	0	1	1	0			1			t_R	212.8	24.69
			1	1	0			-1			t_{exp}	53.94	583.07
0	0	1							1	0	$t_{1(1,0)}$	212.8	23.39
									-1	0	$t_{1(-1,0)}$	53.94	527.29
							1	0	1	1	$t_{2(1,1)}$	143.99	42.93
							1	0	1	-1	$t_{2(1,-1)}$	306.54	14.63
							1	0	0	1	$t_{2(0,1)}$	72.12	348.58
							1	0	0	-1	$t_{2(0,-1)}$	143.97	42.93
			1	1	1	1	1	1	1	1	t_p	384.02	651.04
											Optimum		

Proposed Estimators in Two Phase Sampling

When $\bar{X}$ is unknown, it is sometimes estimated from a preliminary large sample of size n' on which only the characteristic x is measured (for details see Singh et al., 2007). Then, a second phase sample of size n ($n < n'$) is drawn on which both y and x characteristics are measured. Let $\bar{x} = \frac{1}{n'} \sum_{i=1}^{n'} x_i$ denote the sample mean of x based on first phase sample of size n' , $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, be the sample means of y and x respectively based on second phase of size n . In two-phase sampling the estimator t_p will take the following form

$$t_{pd} = \sum_{i=0}^3 h_i t_{id} \in H \quad (20)$$

Such that,

$$\sum_{i=0}^3 h_i = 1 \text{ and } h_i \in R \quad (21)$$

ALMOST UNBIASED ESTIMATOR OF POPULATION PARAMETER

where,

$$t_{0d} = \bar{y}, t_{1d} = \bar{y} \left(\frac{K_1 \bar{x}' + K_1 K_2}{K_1 \bar{x} + K_2 K_3} \right)^m, t_{2d} = \bar{y} \left\{ 2 - \left(\frac{\bar{x}}{\bar{x}'} \right)^q \exp \left[\gamma \left(\frac{(K_4 \bar{x}' + K_5) - (K_4 \bar{x} + K_5)}{(K_4 \bar{x}' + K_5) + (K_4 \bar{x} + K_5)} \right) \right] \right\}$$

The bias and MSE expressions of the estimator t_{1d} and t_{2d} up to the first order of approximation are, respectively, given by

$$B(t_{1d}) = \bar{Y} \left[\frac{m(m-1)R_1^2 f_2 C_x^2}{2} + \frac{m(m+1)R_1^2 f_1 C_x^2}{2} - m^2 R_1^2 f_2 C_x^2 + m R_1 f_3 K_x C_x^2 \right] \quad (22)$$

$$\text{MSE}(t_{1d}) = \bar{Y}^2 \left[f_1 C_y^2 + m^2 R_1^2 f_3 C_x^2 - 2m R_1 K_x f_3 C_x^2 \right] \quad (23)$$

$$B(t_{2d}) = \bar{Y} \left[-\frac{q(q-1)f_1 C_x^2}{2} + \frac{q(q+1)f_2 C_x^2}{2} + q f_2 K_x C_x^2 + q^2 f_2 C_x^2 + f_3 \gamma R_2 K_x C_x^2 + f_3 \gamma R_2 q C_x^2 \right] \quad (24)$$

$$\text{MSE}(t_{2d}) = \bar{Y}^2 \left[f_1 C_y^2 + L_1^2 f_3 C_x^2 \right] \quad (25)$$

where

$$\left. \begin{aligned} R_1 &= \frac{K_1 \bar{X}}{K_1 \bar{X} + K_2 K_3} \\ R_2 &= \frac{K_4 \bar{X}}{2[K_4 \bar{X} + K_5]} \\ L_1 &= q - \gamma A_2 \end{aligned} \right\} \quad (26)$$

Expressing (20) in terms of e 's,

$$t_{pd} = \bar{Y} \left[\begin{aligned} &1 + e_0 + w_1 \left(\frac{m(m+1)R_1^2 e_1^2}{2} - mR_1 e_1 - mR_1 e_0 e_1 + mR_1 e_0 e_1' + \frac{m(m-1)R_1^2 e_1'^2}{2} + mR_1 e_0 e_1' \right) \\ &+ w_2 \left(-qe_1 - \frac{q(q-1)e_1^2}{2} + qe_1' + q^2 e_1 e_1' - \frac{q(q+1)e_1'^2}{2} - \gamma R_2 (e_1' - e_1) + \gamma R_2 (e_0 e_1 - e_0 e_1') - qe_0 e_1 \right) \end{aligned} \right]$$

Subtracting $\bar{Y}$ from both sides of equation (22) and then taking expectation of both sides, the bias of the estimator t_{pd} is obtained up to the first order of approximation, as

$$B(t_{pd}) = \bar{Y} [B(t_{1d}) + B(t_{2d})] \quad (27)$$

also,

$$(t_{pd} - \bar{Y}) = \bar{Y} [e_0 + w_1 [mR_1 e_1' - mR_1 e_1] + w_2 (-qe_1 + qe_1' - \gamma R_2 e_1' + \gamma R_2 e_1)] \quad (28)$$

Squaring both sides of (28) and then taking expectation, the MSE of the estimator t_{pd} is obtained up to the first order of approximation, as

$$\text{MSE}(t_{pd}) = \bar{Y}^2 f_1 C_y^2 + L_2^2 f_3 C_x^2 - 2L_2 f_3 K_x C_x^2 \quad (29)$$

It is a minimum when

$$L_2 = K_x \quad (30)$$

where

$$L_2 = h_1 m R_1 + h_2 (q - \gamma R_2) \quad (31)$$

Putting the value of $L_2 = K_x$ in (28), the optimum value of estimator as t_{pd} (optimum) is obtained. Thus, the minimum MSE of t_{pd} is given by

$$\min. \text{MSE}(t_{pd}) = \bar{Y}^2 C_y^2 (f_1 - f_3 \rho_{yx}^2), \quad (32)$$

which is same as that of traditional linear regression estimator.

From (21) and (31), there are two equations in three unknowns. It is not possible to find the unique values for h_i 's, $i = 0, 1, 2$. In order to get unique values of h_i 's, impose the linear restriction

$$\sum_{i=0}^2 h_i B(t_i) = 0 \quad (33)$$

where $B(t_i)$ denotes the bias in the i^{th} estimator.

Equations (21), (31) and (33) can be written in the matrix form as

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & mR_1 & q - \gamma R_2 \\ 0 & B(t_{1d}) & B(t_{2d}) \end{bmatrix} \begin{bmatrix} h_0 \\ h_1 \\ h_2 \end{bmatrix} = \begin{bmatrix} 1 \\ K_x \\ 0 \end{bmatrix} \quad (34)$$

Using (34), we get the unique values of h_i 's, $i = 0, 1, 2$ as

$$\left. \begin{aligned} h_0 &= 1 - h_1 - h_2 \\ h_1 &= \frac{k_x}{mR_1} - \frac{N_1 K_x (q - \gamma R_2)}{N_1 q - mR_1 N_2 - N_1 \gamma R_2} \\ h_2 &= \frac{K_x N_1}{[N_1 q - mR_1 N_2 - N_1 \gamma R_2]} \end{aligned} \right\}$$

where,

$$N_1 = \frac{m(m-1)R_1^2 f_2 C_x^2}{2} + \frac{m(m+1)R_1^2 f_1 C_x^2}{2} - m^2 R_1^2 f_2 C_x^2 + mR_1 f_3 K_x C_x^2$$

$$N_2 = \left[-\frac{q(q-1)f_1 C_x^2}{2} + \frac{q(q+1)f_2 C_x^2}{2} + qf_2 K_x C_x^2 + q^2 f_2 C_x^2 + f_3 \gamma R_2 K_x C_x^2 + f_3 \gamma R_2 q C_x^2 \right]$$

Use of these h_i 's, $i = 0, 1, 2$ remove the bias up to terms of order $o(n^{-1})$ at (20).

References

- Bahl, S. & Tuteja, R. K. (1991). Ratio and product type exponential estimator. *Journal of Information and Optimization Sciences*, 12(1), 159-163. doi:10.1080/02522667.1991.10699058
- Kadilar, C. and Cingi, H. (2006, March). New ratio estimators using correlation coefficient. *InterStat*, 4, 1-11.
- Khoshnevisan, M. Singh, R., Chauhan, P., Sawan, N., & Smarandache, F. (2007). A general family of estimators for estimating population mean using known value of some population parameter(s). *Far East Journal of Theoretical Statistics*, 22(2), 181-191.
- Sahai A., & Ray, S. K. (1980). An efficient estimator using auxiliary information. *Metrika* 27(4), 271–275.
- Sen, A.R. (1978). Estimation of the population mean when the coefficient of variation is known. *Communications in Statistics – Theory and Methods*, 7(7), 657-672. doi:10.1080/03610927808827656
- Singh, H. P. & Kakran, M. S. (1993). A modified ratio estimator using coefficient of variation of auxiliary character. Unpublished.
- Singh, H. P. & Solanki, R. S. (2012). Improved estimation of population mean in simple random sampling using information on auxiliary attribute. *Applied Mathematics and Computation* 218, 7798–7812. doi:10.1016/j.amc.2012.01.047
- Singh, H. P. & Tailor, R. (2003). Use of known correlation coefficient in estimating the finite population mean. *Statistics in Transition*, 6(4), 555-560.
- Singh, J. Pandey, B.N. & Hirano, K. (1973). On the utilization of a known coefficient of kurtosis in the estimation procedure of variance. *Annals of the Institute of Statistical Mathematics*, 25(1), 51-55.
- Singh, R., Cauhan, P., Sawan, N. & Smarandache, F. (2007). Auxiliary information and a priori values in construction of improved estimators. *Renaissance High Press*.
- Singh, R., Kumar, M., & Smarandache, F. (2008). Almost unbiased estimator for estimating population mean using known value of some population parameter(s). *Pakistan Journal of Statistics and Operation Research*, 4(2), 63-76. doi:10.18187/pjsor.v4i2.50
- Sisodiya, B. V. S. & Dwivedi, V. K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. *Journal of the Indian society of Agricultural Statistics*, 33(2), 13-18.

ALMOST UNBIASED ESTIMATOR OF POPULATION PARAMETER

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S., & Ashok, C. (1984). *Sampling theory of surveys with applications*. Ames, IA: Iowa State University Press.

Upadhyaya, L. N. & Singh, H. P. (1984). On the estimation of the population mean with known coefficient of variation. *Biometrical Journal*, 26(8), 915-922. doi:[10.1002/bimj.4710260814](https://doi.org/10.1002/bimj.4710260814)

Appendix A.

Some members (ratio-type) of the class t_1 when $w_0 = 0$, $w_1 = 1$, $w_2 = 0$, $\alpha = 1$

K_1	K_3	PRE's $K_2 = 1$	PRE's $K_2 = -1$
1	C_x	212.80	212.82
1	$\beta_2(x)$	212.60	213.02
$\beta_2(x)$	C_x	212.81	212.81
C_x	$\beta_2(x)$	212.71	212.91
1	ρ_{yx}	212.81	212.82
$N\bar{X}$	S_x	212.81	212.81
$N\bar{X}$	f	212.80	212.82
$\beta_2(x)$	K_x	212.60	213.02
N	K_x	212.81	212.81
N	1	212.71	212.91
N	C_x	212.81	212.82
N	ρ_{yx}	212.81	212.81
N	S_x	212.80	212.82
N	f	212.60	213.02
N	$g = 1 - f$	212.81	212.81
N	K_x	212.71	212.91
n	ρ_{yx}	212.81	212.82
n	S_x	212.81	212.81
n	f	212.80	212.82
n	$g = 1 - f$	212.60	213.02
n	K_x	212.81	212.81
$\beta_2(x)$	$\bar{X}$	212.71	212.91
$N\bar{X}$	$\bar{X}$	212.81	212.82
N	$\bar{X}$	212.81	212.81
n	$\bar{X}$	212.81	212.81

Appendix B.

Some members (product-type) of the class t_1 when $w_0 = 0$, $w_1 = 1$, $w_2 = 0$, $\alpha = -1$

K_1	K_3	PRE's $K_2 = 1$	PRE's $K_2 = -1$
1	C_x	550.91	501.92
1	$\beta_2(x)$	646.03	314.33
$\beta_2(x)$	C_x	535.22	519.18
C_x	$\beta_2(x)$	582.35	91.18
1	ρ_{yx}	466.00	579.15
$N\bar{X}$	S_x	528.52	526.06
$N\bar{X}$	f	527.30	527.28
$\beta_2(x)$	K_x	510.01	543.74
N	K_x	527.15	527.43
N	1	530.39	524.16
N	C_x	528.52	526.03
N	ρ_{yx}	524.41	530.15
N	S_x	549.55	503.49
N	f	527.53	527.06
N	$g = 1 - f$	530.16	524.40
N	K_x	524.70	529.87
n	ρ_{yx}	520.05	534.38
n	S_x	579.44	465.58
n	f	527.88	526.71
n	$g = 1 - f$	534.42	520.01
n	K_x	520.77	533.69
$\beta_2(x)$	$\bar{X}$	622.76	146.09
$N\bar{X}$	$\bar{X}$	530.39	524.16
N	$\bar{X}$	580.14	464.59
n	$\bar{X}$	632.80	363.64

Appendix C.

Some members (product-type) of the class t_2 when $w_0 = 0$, $w_1 = 0$, $w_2 = 1$

K_4	K_5	PRE'S
K_1	K_3	$(\beta = -1, \lambda = -1)$
1	C_x	358.00
1	$\beta_2(x)$	423.38
$\beta_2(x)$	C_x	351.94
C_x	$\beta_2(x)$	357.48
1	ρ_{yx}	324.10
$N\bar{X}$	S_x	349.09
$N\bar{X}$	f	348.58
$\beta_2(x)$	K_x	341.45
N	K_x	348.52
N	1	349.89
N	C_x	349.09
N	ρ_{yx}	347.37
N	S_x	358.21
N	f	348.68
N	$g = 1 - f$	349.79
N	K_x	347.49
n	ρ_{yx}	345.56
n	S_x	372.38
n	f	348.82
n	$g = 1 - f$	351.60
n	K_x	345.86
$\beta_2(x)$	$\bar{X}$	345.86
$N\bar{X}$	$\bar{X}$	349.89

In addition to above estimators a large number of estimators can also be generated from the proposed estimators just by putting different values of constants w_i 's, h_i 's $K_1, K_2, K_3, K_4, K_5, \alpha, \beta$ and λ .

Application of Esscher Transformed Laplace Distribution in Microarray Gene Expression Data

Shanmugasundaram Devika
Christian Medical College
Vellore, India

Sebastian George
St. Thomas College
Palai, Kerala, India

Lakshmanan Jeyaseelan
Christian Medical College
Vellore, India

Microarrays allow the study of the expression profile of hundreds to thousands of genes simultaneously. These expressions could be from treated samples and the healthy controls. The Esscher transformed Laplace distribution is used to fit microarray expression data as compared to Normal and Laplace distributions. The Maximum Likelihood Estimation procedure is used to estimate the parameters of the distribution. R codes are developed to implement the estimation procedure. A simulation study is carried out to test the performance of the algorithm. AIC and BIC criterion are used to compare the distributions. It is shown that the fit of the Esscher transformed Laplace distribution is better as compared to Normal and standard Laplace distributions.

Keywords: Esscher transformed Laplace distribution, Normal distribution, Laplace distribution, Microarray gene expression, Maximum Likelihood estimation

---

Introduction

Microarrays allow the researcher to investigate the expressions of thousands of genes simultaneously under various condition of the biological process. These conditions could be samples from cancer tumor and healthy controls. This method measures the intensity of the fluorescence after hybridization and then expression profiles are compared between two different samples of Complementary DNA (cDNA) colored with different dyes, Red (for diseased) and Green (for healthy control). Hence this method allows us to study the relative gene expression in two different samples. The statistical methods that have been developed to analyze the gene expression data over the decades depend heavily on the distribution of the gene expression data.

Ms. Devika is a PhD Research Scholar. Email her at: devika@cmcvellore.ac.in. Dr. George is a Professor in the Department of Statistics. Email him at: sthotom@gmail.com. Dr. Jeyaseelan is a Professor. Email him at: ljey@cmcvellore.ac.in.

The gene expression data, after normalization, usually has a heavier tail as compared to normal distribution. That is, most of the mass at the center with a sharp peak with varying asymmetry. Researchers have used several densities to model gene expression data. Densities of Poisson, exponential, and logarithmic series were used (Kuznetsov, 2001). An error distribution of gene expression datasets was approximated by two distributions by taking log-normal in the bulk of microarray spot intensities and a power law in the tails (Hoyle, Rattray, Jupp, & Brass, 2002). The gene expression was also fitted by using an asymmetric Laplace distribution (Purdom & Holmes, 2005). However, in order to take outliers into account, the Cauchy distribution has been used for estimating gene expressions using data from multiple-laser scans (Khondoker, Glasbey, & Worton, 2006), and the Laplace mixture model was introduced as a long tailed alternative to the normal distribution (Bhowmick, Davison, Goldstein, & Ruffieux, 2006).

Recently, asymmetric type II compound Laplace density (Punathumparambath, Kulathinal, & George, 2012) was introduced for the analysis of gene expression data which was asymmetric version of type II compound Laplace distribution and a generalization of asymmetric Laplace distribution. The four parameter probability distribution provided an additional degree of freedom to capture the characteristic feature of the microarray data. Based on the above review, the microarray data with thousands of genes show asymmetry and most of the mass at the middle as large proportion of genes are not differently expressed. Therefore the log ratio of the intensities have a tendency to cluster around a single point and with the presence of outliers. Hence it may not be appropriate to summarize such pattern with mean, variance, etc.

In the current study, new class of asymmetric Laplace distribution is proposed for the analysis of log ratios of measured gene expression data across genes through Esscher transformation, namely Esscher transformed Laplace (ETL) distribution proposed in George and George (2012). It is a sub-class of one parameter exponential family and an alternative to various types of asymmetric Laplace distributions given in Kotz, Kozubowski, and Podgórski (2001). If all the genes on one array are considered as separate independent observations, the distribution of the log-ratio of the expression values is well approximated by the asymmetric nature of the ETL distribution. Moreover modeling distribution with single parameter would be a feasible approach as compared to distribution such as asymmetric type II compound Laplace distribution with four parameters. This paper presents the analysis of microarray gene expression data using the ETL distribution. The paper is organized as follows: First we describe the overview of ETL distribution, followed by a simulation study. Next Normal, Laplace, and ETL

distributions were fitted to gene expression data and compared. Finally the paper ends with conclusion.

Methods

Overview of Esscher Transformed Laplace Distribution

The ETL distribution was proposed in George and George (2012) and George (2011). A random variable X is said to follow Esscher transformed Laplace distribution with parameter (θ) if its probability distribution function (pdf) is given by

$$f(x|\theta) = \frac{1-\theta^2}{2} \begin{cases} \exp[x(1+\theta)], & x < 0 \\ \exp[-x(1-\theta)], & x \geq 0 \end{cases} \quad (1)$$

where θ is called the Esscher parameter and $\theta \in (-1, 1)$. This pdf can also be expressed as

$$f(x|\theta) = \frac{1-\theta^2}{2} \exp[\theta x - |x|]; -\infty < x < \infty, \theta \in (-1, 1) \quad (2)$$

Thus the ETL distribution is a regular one parameter exponential family and a subclass of the family of asymmetric Laplace (AL) distributions proposed in Kotz et al. (2001). These kinds of distribution are more appropriate for modeling financial datasets as this allows for asymmetry, peakedness and tailed heaviness than normal distribution (George & George, 2013).

The cumulative distribution function (cdf) of the ETL distribution is given by

$$F(x|\theta) = \begin{cases} \frac{(1-\theta)}{2} \exp[x(1+\theta)], & x < 0 \\ \frac{1-\theta}{2} + \frac{1+\theta}{2} (1 - \exp[-x(1-\theta)]), & x \geq 0 \end{cases} \quad (3)$$

for $\theta \in (-1, 1)$. When $\theta = 0$, we get the classical Standard Laplace (0, 1) distribution.

Figure 1 represents the densities of the ETL distribution. When $\theta \in (-1, 0)$ the distribution is left skewed and $\theta \in (0, 1)$ the distribution is right skewed. From

Figure 1 we can see that the ETL distribution has heavier tails than the normal distribution, meaning that there is more probability of extreme values than under a normal distribution. In addition, the ETL distribution concentrates more probability in the center than a normal distribution. It is also clear from Figure 1 that the shape of the ETL distribution is nearly similar to the AL distribution but the later does not belong to one parametric exponential family whereas the former does.

The characteristic function of the AL (μ, σ) with parameters $\mu \in \mathbb{R}$ and $\sigma \geq 0$ and ETL (θ) distributions are given by

$$\varphi_X(t) = (1 + \sigma^2 t^2 - i\mu t)^{-1}$$

and

$$\varphi_X(t) = \left(1 + \frac{t^2}{1 - \theta^2} - \frac{2it\theta}{1 - \theta^2} \right)^{-1}$$

Figure 1. Densities of the Esscher transformed Laplace distribution for various choices of parameter θ .

Hence ETL (θ) is a special case of the AL (μ, σ) distribution with $\mu = 2\theta/(1 - \theta^2)$ and $\sigma^2 = 1/(1 - \theta^2)$. The mean $E(x)$ and variance $\text{Var}(x)$ of the ETL distribution are given by

$$E(x) = \frac{2\theta}{1-\theta^2}; \quad \text{Var}(x) = \frac{2(1+\theta^2)}{(1-\theta^2)}$$

The α^{th} quantile of the ETL (θ) distribution for simulation purpose in the later section is given by

$$q_\alpha = \begin{cases} \frac{1}{1+\theta} \log \left[\frac{2}{1-\theta} \alpha \right], & \alpha \in \left(0, \frac{1-\theta}{2} \right) \\ \frac{1}{1+\theta} \log \left[\frac{2}{1+\theta} (1-\alpha) \right], & \alpha \in \left(\frac{1-\theta}{2}, 1 \right) \end{cases} \quad (4)$$

The parameter of the ETL distribution can be obtained either by the method of maximum likelihood (MLE) or by the method of moments. Let $x_1, x_2, \dots, x_n$ be an independent identically distributed (i.i.d) random variable from the ETL (θ) distribution with density from equations (1) or (2). The likelihood function is then written as

$$\log L(X : \theta) = n \log \left(\frac{1-\theta^2}{2} \right) + \theta \sum_{i=1}^n x_i - \sum_{i=1}^n |x_i|$$

and the first derivative with respect to the parameter θ is

$$\frac{\partial \log L}{\partial \theta} = \frac{-2n\theta}{1-\theta^2} - \sum_{i=1}^n x_i$$

The MLE of parameter θ is obtained by solving the score function

$$\frac{\partial \log L}{\partial \theta} = 0$$

so that

$$\hat{\theta} = \frac{-1 \pm \sqrt{1 + \bar{x}^2}}{\bar{x}}$$

provided that $\theta \in (-1, 1)$.

By introducing the location parameter (μ) and scale parameter (σ) in the ETL distribution, the pdf and cdf of the ETL (θ, μ, σ) distribution is given as follows:

$$f(x|\theta, \mu, \sigma) = \begin{cases} \frac{(1-\theta^2)}{2\sigma} \exp\left[\left(\frac{x-\mu}{\sigma}\right)(1+\theta)\right], & x < \mu \\ \frac{(1-\theta^2)}{2\sigma} \exp\left[\left(\frac{\mu-x}{\sigma}\right)(1-\theta)\right], & x \geq \mu \end{cases} \quad (5)$$

and

$$F(x|\theta, \mu, \sigma) = \begin{cases} \frac{(1-\theta)}{2} \exp\left[\left(\frac{x-\mu}{\sigma}\right)(1+\theta)\right], & x < \mu \\ 1 - \frac{(1+\theta)}{2} \exp\left[\left(\frac{\mu-x}{\sigma}\right)(1-\theta)\right], & x \geq \mu \end{cases} \quad (6)$$

where $\theta \in (-1, 1)$, $\mu \in \mathbb{R}$, and $\sigma > 0$.

The mean $E(x)$ and variance $\text{Var}(x)$ of the ETL with location μ and scale parameter σ are given by

$$E(x) = \mu + \frac{2\theta\sigma}{1-\theta^2}$$

$$\text{Var}(x) = \frac{2\sigma^2(1+\theta^2)}{(1-\theta^2)^2}$$

The α^{th} quantile of the ETL (θ, μ, σ) distribution is

$$q_a = \begin{cases} \mu + \frac{\sigma}{1+\theta} \log\left(\frac{2}{1-\theta} q\right), & q \in \left(0, \frac{1-\theta}{2}\right) \\ \mu - \frac{\sigma}{1-\theta} \log\left(\frac{2}{1+\theta} (1-q)\right), & q \in \left(\frac{1-\theta}{2}, 1\right) \end{cases} \quad (7)$$

The parameters θ , μ , and σ of the ETL distribution were obtained by maximization of the likelihood function in R software (R Development Core Team, 2014) using optim function with BFGS (Broyden, Fletcher, Goldfarb, and Shanno) algorithm. The standard error (SE) of the respective parameters were obtained by inverting the Fisher information matrix at the maximum likelihood estimates. As this was a methodological study which used open source data, IRB clearance was not necessary.

Data Simulation

A simulation experiment is executed to study the functioning of the estimation algorithm for various arbitrary values of the parameters of the ETL (θ, μ, σ) distribution. We created 1000 datasets each with sample of size $n = 2000$ from the ETL distribution by fixing the Esscher parameter $\theta = (-0.5, 0, 0.5)$, location parameter $\mu = (-0.5, -0.2, 0.3, 0.9)$, and scale parameter $\sigma = (0.5, 0.75, 1, 1.5)$ by using an inverse transform sampling procedure. Then the maximum likelihood estimates of the parameters are obtained as mentioned above by using R statistical software. Table 1 represents the results of the simulation study performed by using 1000 different data sets. It is apparent that the estimation procedure works well for different choices of parameters and the sample standard deviation are in accordance with the asymptotic standard error obtained using maximum likelihood estimate. However the difference increases with increase in the σ values. We also checked the convergence of the estimation procedure for various choices of parameter values with different initials and the algorithm works satisfactorily well for several alternatives.

Results

Analysis of Microarray gene expression data

The ETL distribution was applied to three different microarray datasets (Swirl, E. coli, and Tumor) from published microarray experiments. The first data set Swirl

zebrafish experiment is included as part of the marray package in R software (Dudoit & Yang, 2002). This data is provided by Katrin Wuennenberg-Stapleton from the Ngai Lab at UC Berkeley (2001). Swirl is a point mutant in the vertebrates. In order to access the mutational status, zebrafish was taken as a model organism. The aim of the experiment was to find genes which were differentially expressed between mutant and wild type zebrafish. The cDNA from wild type mutant was labelled using Cy3 dyes and the swirl mutant with Cy5. There were totally four replicates (Swirl.1, ..., Swirl.4) and the target cDNA was hybridized to microarrays containing 8,448 probes, including 768 control spots. The raw dataset was first log transformed to base 2 and normalized using a print tip group Lowess smoothing technique (locally weighted linear regression method) (Cleveland & Devlin, 1988) and with quantile normalization procedure. This method is widely used in microarray experiments as this removes the intensity dependence in $\log_2(R_i/G_i)$ values, where R_i is the red dye intensity (Cy3) and G_i (Cy5) is the green dye intensity for the i^{th} gene (Yang et al., 2002). The same dataset was used to fit asymmetry Laplace distribution in Purdom and Holmes (2005).

The next dataset, E. coli, was a two channel microarray experiment conducted to compare gene expression profiles of wild strain with mutant strain and was provided by Bernstein, Lin, Cohen, and Lin-Chao (2004). The dataset contained information on 5128 genes with six arrays. mRNA extracted from wild strain was labeled with Cy5 (Green) and the mutant strain with Cy3 (Red). The E. coli data was also normalized using Lowess technique and the quantile normalization procedure and then the log differences was taken as gene expression measurement. The third dataset Tumor microarray experiment was carried on to compare the functioning of gene expression of ovarian tumor cells as compared to normal cells. This study involved six samples from normal cells and six from ovarian tumor cells on 34,742 genes. We transformed the data using log function with base 2 and then we used Lowess and quantile normalization procedure as earlier.

Gaussian, Laplace, and ETL distributions were fitted to log transformed normalized gene expression measurements $\log_2(R_i/G_i)$ for the three datasets. The parameters of the Gaussian (μ, σ^2), Laplace (μ, σ), and ETL (θ, μ, σ) distributions were estimated using maximum likelihood estimation method and their corresponding standard errors. In Table 2, results for two arrays from each dataset are presented, and the rest are given in the supplementary Table 4.

ETL DISTRIBUTION IN MICROARRAY DATA

Table 1. Simulation study – maximum likelihood estimates of θ , μ , and σ for various choices of parameters

θ	σ	μ	$\hat{\theta}$	$\hat{\sigma}$	$\hat{\mu}$	SE($\hat{\theta}$)	SD($\hat{\theta}$)	SE($\hat{\sigma}$)	SD($\hat{\sigma}$)	SE($\hat{\mu}$)	SD($\hat{\mu}$)
-0.5	0.50	-0.5	-0.4999	0.5001	-0.5001	0.0177	0.0199	0.0163	0.0171	0.0147	0.0190
	0.75	-0.2	-0.4999	0.7500	-0.2000	0.0173	0.0199	0.0241	0.0257	0.0204	0.0286
	1.00	0.3	-0.4999	1.0001	0.3000	0.0170	0.0199	0.0319	0.0342	0.0259	0.0382
	1.50	0.9	-0.5000	1.5000	0.9000	0.0166	0.0199	0.0472	0.0514	0.0358	0.0572
0.0	0.50	-0.5	0.0003	0.5000	-0.5002	0.0210	0.0227	0.0112	0.0110	0.0135	0.0168
	0.75	-0.2	0.0003	0.7500	-0.2004	0.0203	0.0227	0.0168	0.0165	0.0187	0.0252
	1.00	0.3	0.0003	1.0000	0.2995	0.0199	0.0228	0.0224	0.0220	0.0235	0.0337
	1.50	0.9	0.0003	1.5001	0.8992	0.0197	0.0228	0.0336	0.0329	0.0337	0.0505
0.5	0.50	-0.5	0.5015	0.4987	-0.5017	0.0180	0.0199	0.0165	0.0174	0.0150	0.0199
	0.75	-0.2	0.5015	0.7481	-0.2025	0.0172	0.0200	0.0241	0.0261	0.0203	0.0300
	1.00	0.3	0.5015	0.9974	0.2966	0.0171	0.0199	0.0320	0.0347	0.0261	0.0399
	1.50	0.9	0.5016	1.4961	0.8949	0.0166	0.0200	0.0472	0.0521	0.0362	0.0599

Figures 2-3 represent the box plots of intensities of Swirl, E. coli, and Tumor datasets before and after normalization. It is clear from Figures 2-3 that, after normalization, each distribution of the gene expression has a similar shape and exhibits heavier tails with a certain degree of asymmetry as compared to a Gaussian distribution. The left side of Figures 4-9 and supplementary Figures 10-19 shows the histogram super imposed with ETL (θ, μ, σ), Laplace (μ, σ) and Gaussian (μ, σ^2) distributions, where the parameters of these distributions were obtained by the maximum likelihood estimation procedure. By comparing these densities, ETL (θ, μ, σ) captures the asymmetric nature of the data with peaked concentration in the middle and heavy tail.

It can be seen from Table 2 that the Esscher parameter (θ) for arrays Swirl.1 and Swirl.3 are greater than 0 (right skewed) and for all the other arrays the parameter (θ) is smaller than 0 (left skewed). Though the level of skewness in all the arrays of the datasets is not very large, they are different from 0. It is also noted that the maximum likelihood estimate of parameter σ of the ETL and Laplace distributions are approximately equal.

Figure 2. Boxplot of intensities from Swirl zebrafish microarray experiment, before and after normalization.

Figure 3. Boxplot of intensities of Red and Green arrays of Ecoli and Tumor microarray experiments, before and after normalization.

ETL DISTRIBUTION IN MICROARRAY DATA

Table 2. Microarray data analysis – maximum likelihood estimates and the asymptotic standard error for Esscher transformed Laplace, Laplace, and Normal distributions.

	Swirl.1	Swirl.3	Ecoli.1	Ecoli.2	Tumor.3	Tumor.5
Esscher						
θ	0.24(0.0128)	0.23(0.0111)	-0.090(0.0188)	-0.160(0.0159)	-0.080(0.0063)	-0.080(0.0064)
σ	0.26(0.0034)	0.30(0.0036)	0.330(0.0047)	0.430(0.0065)	0.710(0.0039)	0.660(0.0036)
μ	-0.09(0.0058)	-0.10(0.0052)	0.060(0.0106)	0.140(0.0112)	0.110(0.0072)	0.110(0.0069)
Laplace						
μ	-0.01(0.0035)	-0.01(0.0038)	0.020(0.0060)	0.050(0.0082)	0.040(0.0047)	0.040(0.0043)
σ	0.29(0.0031)	0.32(0.0035)	0.330(0.0046)	0.450(0.0063)	0.710(0.0038)	0.660(0.0036)
Gaussian						
μ	0.05(0.0052)	0.04(0.0047)	0.002(0.0068)	0.002(0.0092)	0.005(0.0054)	0.005(0.0051)
σ	0.23(0.0035)	0.19(0.0029)	0.240(0.0047)	0.430(0.0085)	1.030(0.0078)	0.890(0.0068)

One of the graphical procedures to compare the probability distribution Quantile-Quantile plot (Q-Q plot) is shown in the right side of Figures 4-9 and supplementary Figures 10-19. This is obtained by plotting the theoretical quantiles against sample quantiles. This plot is more useful as this better emphasizes the fit of the distributions in the tail region. It is indicated in Figures 4-9 that the ETL (θ, μ, σ) distribution fits to the data well as compared to other two distributions, especially when (θ) is significantly greater than 0 (right skewed) for Swirl.1 and 3 and smaller than 0 (left skewed) for all the other arrays. The supplementary Figures 10-19 indicate that, when $\theta \approx 0$, the performance of both the Laplace and ETL distributions are almost similar but still better than Gaussian distribution. Other than with few outliers, the fit of the ETL distribution is greatly improved as compared to the other distributions considered, though all the three seem to describe the middle region of the data rather similarly.

A numerical evaluation of model comparison was done by using Akaike's Information Criterion (AIC) (Akaike, 1998) and Bayesian Information Criterion (BIC) (Schwarz, 1978) as the later take into account of the sample size. The formula for AIC and BIC are given by

$$\text{AIC} = -2 \log \left(L_g \left(\hat{\theta} \mid x_1, \dots, x_n \right) \right) + 2K$$

and

$$\text{BIC} = -2\log\left(L_g\left(\hat{\theta} \mid x_1, \dots, x_n\right)\right) + K\log(n)$$

where K is the number of parameters being estimated, L is the likelihood function of the model g , $\hat{\theta}$ is the maximum likelihood estimate of the parameters of model g , and n is the sample size. Given the different models, the one with smaller AIC/BIC fits the data better than the one with the larger AIC/BIC, where the conclusion from AIC and BIC goes hand in hand in most of the cases. AIC and BIC values of the three distributions, ETL (θ, μ, σ) , Laplace (μ, σ) , and Gaussian (μ, σ^2) are given in Table 3 and supplementary Table 5. The ETL (θ, μ, σ) distribution had a lower AIC/BIC values for all the sample arrays shown in Table 3. Hence the ETL distribution shows an improvement in the model fit as compared to other distributions. However, when there is an absence of asymmetry ($\theta \approx 0$) the values of AIC/BIC for the ETL distribution are nearly equal to the Laplace distribution. This feature has been seen in the arrays of Swirl.2, Ecoli.4, Ecoli.5, Ecoli.6 and Tumor.2 in supplementary Table 5, which shows a similar performance of ETL and Laplace distributions.

Figure 4. Left: Histogram of Swirl.1 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

ETL DISTRIBUTION IN MICROARRAY DATA

Figure 5. Left: Histogram of Swirl.3 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

Figure 6. Left: Histogram of Ecoli.1 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

Figure 7. Left: Histogram of Ecoli.2 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

Figure 8. Left: Histogram of Tumor.3 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

ETL DISTRIBUTION IN MICROARRAY DATA

Figure 9. Left: Histogram of Tumor.5 superimposed with Esscher transformed Laplace (red line), Laplace (blue dotted), and Normal (green dashed) distributions. Right: Q-Q plot of Esscher transformed Laplace (red), Laplace (blue), and Normal (green) distributions.

Table 3. Comparison of AIC and BIC of Esscher transformed Laplace, Laplace, and Normal distributions.

	Swirl.1		Swirl.3		Ecoli.1		Ecoli.2		Tumor.3		Tumor.5	
	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
Esscher	7125	7146	8942	8963	6023	6043	9084	9104	94157	94183	89044	89069
Laplace	7549	7563	9245	9259	6045	6058	9167	9180	94301	94318	89200	89217
Gaussian	11406	11420	9855	9869	7234	7247	10248	10261	99634	99651	94568	94585

Conclusion

In the two channel microarray experiments, for which the ETL distribution was fitted, gave a reasonable fit to the gene expression data and greatly improved upon the normal distribution and as an alternative to Laplace distribution. The ETL (θ, μ, σ) can be a better model for gene expression data as they are asymmetric, heavy tailed, and with bulk mass in the middle of the distribution and which does not follow any of the classical symmetric distributions such as Normal, Laplace etc., Esscher transformed Laplace distribution is simple to use distribution which belongs to regular exponential family captures all the features as mentioned above

of the gene expression measurement. In this distribution, the asymmetry is determined by using Esscher parameter (θ) along with the location (μ) and scale (σ) parameters. This distribution is more flexible and belongs to the special case of AL distribution and is also easily tractable for statistical inference. Simulating observations from the ETL distribution is also possible by inverting the cumulative distribution function.

The microarray gene expression data has been modeled using different densities by several authors. AL distribution was introduced in Purdom and Holmes (2005) in the analysis of gene expression data to capture the peak at the center as well as the asymmetry in the distribution. The Laplace mixture model as a long tailed alternative to the normal distribution in identifying differentially expressed genes in microarray experiments was introduced in Bhowmick et al. (2006). The Cauchy distribution was applied in Khondoker et al. (2006) in modeling microarray experiments which can estimate gene expressions by taking the outliers into account. Asymmetric type II compound Laplace distribution in the analysis of microarray gene expression data was introduced in (Punathumparambath et al., 2012). The same author has proposed a family of skew-slash distributions generated by normal kernel (Punathumparambath, 2011), two compound mixture Gaussian models (Punathumparambath, George, & V. M., 2011), skew-slash distributions generated by the Cauchy kernel (Punathumparambath, 2013), skew-slash t and skew-slash Cauchy distributions (Punathumparambath, 2012b), and asymmetric slash Laplace distribution (Punathumparambath, 2012a) for modeling gene expression data.

The ETL distribution was used in modeling microarray data as an alternative to normal and Laplace distributions. From Figures 4-9 and supplementary Figures. 10-19, we can see that the ETL distribution fits the tail region better as compared to other two distributions. This is also evident in the reduction in AIC/BIC values for the ETL distribution as compared to the normal and Laplace distributions. The ETL belongs to exponential family of distributions and is also a generalization of the AL distribution. The main motive of applying different distributions to microarray gene expression data is to capture the asymmetry and peakedness because a large proportion of genes are not differentially expressed, the log ratio of the intensities have tendency to cluster around a single point, and the presence of outliers (Punathumparambath et al., 2012). This distribution is already been applied in George and George (2013) to financial data modeling and web server data, and it was shown that the model fit was better as compared to other distributions.

References

- Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In E. Parzen, K. Tanabe, & G. Kitagawa (Eds.), *Selected Papers of Hirotugu Akaike* (199-213). New York: Springer New York.
- Bernstein, J. A., Lin, P.-H., Cohen, S. N., & Lin-Chao, S. (2004). Global analysis of *Escherichia coli* RNA degradosome function using DNA microarrays. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9), 2758-2763. doi: [10.1073/pnas.0308747101](https://doi.org/10.1073/pnas.0308747101)
- Bhowmick, D., Davison, A. C., Goldstein, D. R., & Ruffieux, Y. (2006). A Laplace mixture model for identification of differential expression in microarray experiments. *Biostatistics*, 7(4), 630-641. doi: [10.1093/biostatistics/kxj032](https://doi.org/10.1093/biostatistics/kxj032)
- Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596-610. doi: [10.2307/2289282](https://doi.org/10.2307/2289282)
- Dudoit, S. & Yang, Y. H. (2002). *marrayClasses package: Classes and methods for cDNA microarray data*. Retrieved from <http://www.bioconductor.org/>
- George, D. (2011). *A class of heavy-tailed distributions and their applications* (Unpublished doctoral thesis). Mahatma Gandhi University, Kottayam, Kerala, India.
- George, D., & George, S. (2011). Application of Esscher transformed Laplace distribution in web server data. *Journal of Digital Information Management*, 9(1), 19-26.
- George, D., & George, S. (2013). Marshall–Olkin Esscher transformed Laplace distribution and processes. *Brazilian Journal of Probability and Statistics*, 27(2), 162-184. doi: [10.1214/11-BJPS163](https://doi.org/10.1214/11-BJPS163)
- George S, & George, D. (2012). Esscher transformed Laplace distribution and its applications. *Journal of Probability and Statistical Science*, 10(2), 135-152.
- Hoyle, D. C., Rattray, M., Jupp, R., & Brass, A. (2002). Making sense of microarray data distributions. *Bioinformatics*, 18(4), 576-584. doi: [10.1093/bioinformatics/18.4.576](https://doi.org/10.1093/bioinformatics/18.4.576)
- Khondoker, M. R., Glasbey, C. A., & Worton, B. J. (2006). Statistical estimation of gene expression using multiple laser scans of microarrays. *Bioinformatics*, 22(2), 215-219. doi: [10.1093/bioinformatics/bti790](https://doi.org/10.1093/bioinformatics/bti790)

- Kotz, S., Kozubowski, T. J., & Podgórski, K. (2001). *The Laplace distribution and generalizations: A revisit with applications to communications, economics, engineering, and finance*. Boston: Birkhäuser.
- Kuznetsov, V. A. (2001). Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. *EURASIP Journal on Advances in Signal Processing*, 2001(4), 285-296. doi: 10.1155/S1110865701000294
- Punathumparambath, B. (2011). A new family of skewed slash distributions generated by the normal kernel. *Statistica*, 71(3), 345-353. doi: 10.6092/issn.1973-2201/3618
- Punathumparambath, B. (2012a). The multivariate asymmetric slash Laplace distribution and its applications. *Statistica*, 72(2), 235-249. doi: 10.6092/issn.1973-2201/3645
- Punathumparambath, B. (2012b). The multivariate skew-slash t and skew-slash Cauchy distributions. *Model Assisted Statistics and Applications*, 7(1), 33-40. doi: 10.3233/MAS-2011-0199
- Punathumparambath, B. (2013). A new family of skewed slash distributions generated by the Cauchy kernel. *Communications in Statistics - Theory and Methods*, 42(13), 2351-2361. doi: 10.1080/03610926.2011.599508
- Punathumparambath, B., George, S., & V. M., K. (2011). Statistical techniques for microarray technology. *Journal of Informatics and Mathematical Sciences*, 3(3), 257-275.
- Punathumparambath, B., Kulathinal, S., & George, S. (2012). Asymmetric type II compound Laplace distribution and its application to microarray gene expression. *Computational Statistics & Data Analysis*, 56(6), 1396-1404. doi: 10.1016/j.csda.2011.10.026
- Purdom, E., & Holmes, S. P. (2005). Error distribution for gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 4(1), . doi: 10.2202/1544-6115.1070
- R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461-464. doi: 10.1214/aos/1176344136
- University of California, Berkeley (2001). *Swirl experimental data* [Data set]. Provided by the Ngai Lab.
- Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., & Speed, T. P. (2002). Normalization for cDNA microarray data: A robust composite method

addressing single and multiple slide systematic variation. *Nucleic Acids Research*, 30(4), e15. doi: [10.1093/nar/30.4.e15](https://doi.org/10.1093/nar/30.4.e15)

New Procedures of Estimating Proportion and Sensitivity Using Randomized Response in a Dichotomous Finite Population

Tanveer A. Tarray
Vikram University
Ujjain, M. P., India

Housila P. Singh
Vikram University
Ujjain, M. P., India

The problem of estimating the population proportion possessing a sensitive attribute using simple random sampling with replacement (SRSWR) is advocated. Two new procedures are proposed. The suggested models are more efficient than the Huang (2004) randomized response technique under some realistic conditions. Numerical and graphic illustrations are given.

Keywords: Randomized response technique, direct response, estimation of proportion, privacy of respondents, sensitive characteristics

---

Introduction

Socioeconomic investigations often relate to certain personal features that people desire to hide from others in comprehensive inquiries, detailed questionnaires include numerous items. Direct questioning of respondents about them is likely to result either in non-response or in a deliberately incorrect answer. Social stigma and fear of reprisals often lead respondents to give biased, misleading, or even erroneous responses when approached with a direct response (DR) survey method. Even for the reason of merely being unwilling to reveal secrets to strangers, many individuals attempt to avoid certain questions put to them by interviewers.

Consider a dichotomous population in which every person belongs to either a sensitive group A or to the non-sensitive complement A^c . The aim is to estimate π , the population proportion of individuals who are members of A . To do so, a simple random sample of size n is drawn from the population with replacement. Let T be the probability that the respondents belong to A report the truth. The respondents

---

Dr. Tarray is affiliated with the Department of Computer Science and Engineering. Email him at: tanveerstat@gmail.com. Dr. Singh is affiliated with the School of Studies in Statistics.

belonging to the non-sensitive group A^c have no reason to tell a lie. For a DR survey of size n , the interviewee is asked if they are a member of A . Then we have a direct estimator

$$\hat{\pi}_D = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

with mean square error (MSE) given by

$$\text{MSE}(\hat{\pi}_D) = \frac{\theta_D(1-\theta_D)}{n} + \pi^2(1-T)^2 \quad (2)$$

where $X_i = 1(0)$ if the i^{th} interviewee responds Yes (No) and $\theta_D = \pi T$.

To procure reliable sample data for the population proportion of the respondents belonging to the sensitive group A , Warner (1965) proposed an ingenious procedure called Warner's randomized response technique. This pioneering work led to modification and developments in several directions; for instance, see Fox and Tracy (1986), Mangat and Singh (1990), Mangat (1994), Mahmood, Singh, and Horn (1998), Chua and Tsui (2000), Sing, Singh, and Mangat (2000), Chang and Huang (2001), Huang (2004), Chang, Wang, and Huang (2004a, b) and Singh and Tarray (2012, 2013a, b, c, 2014a, b, c, d, e).

Huang (2004) pointed out there are many variants of the randomized response technique in the literature, but most do not dwell on the fundamental question: whether or not the issues considered in the survey should be regarded as sensitive, meaning that there is a need for a randomized response procedure rather than a direct response procedure. In general, the probability T is a measure instrument of the sensitivity (see Huang, 2004). It has a primary use in appraising the efficiency of different survey plans. One may use a simple formula for ascertaining whether a randomized response technique is beneficial in efficiency relative to a DR scheme. However, the probability T is unknown in actual practice. To overcome such a difficulty, Chang and Huang (2001), Huang (2004), and Chang et al. (2004a) have suggested alternative survey strategies which make it possible to estimate the unknown parameters π and T simultaneously. Two alternatives to Huang's (2004) randomized response model, based on Singh (1993) models, are proposed.

A Brief Review of Randomized Response Models

Warner's Models

In order to improve respondent cooperation and to encourage honest response, Warner (1965) proposed the following procedure, known as a randomized response technique (RRT). Instead of a DR procedure, a randomization device is used to gather sample information consisting of one of two statements:

- (i) "I am a member of group A"
- (ii) "I am not a member of group A"

with probabilities P and $(1 - P)$ respectively. Following this device, the respondent selects a statement unobserved by the interviewer, and then simply gives a "Yes" or "No" answer in a random sample of n respondents. By the method of moments, Warner obtained an unbiased estimator of the population proportion π possessing the sensitive attribute A:

$$\hat{\pi}_w = \frac{\bar{Y} - (1 - P)}{2P - 1}, \quad P \neq \frac{1}{2},$$

where $Y_i = 1(0)$ if the i^{th} respondent answers Yes (No) and $\bar{Y} = \frac{1}{n} \sum_{i=1}^n y_i$. The variance of $\hat{\pi}_w$ is given by

$$\begin{aligned} V(\hat{\pi}_w) &= \frac{\theta_w (1 - \theta_w)}{n(2P - 1)^2}, \\ &= \frac{\pi(1 - \pi)}{n} + \frac{P(1 - P)}{n(2P - 1)^2}, \end{aligned} \quad (3)$$

where $\theta_w = \pi P + (1 - \pi)(1 - P)$.

Singh Models

Singh (1993) developed two randomized response techniques named RRT1 and RRT2 which are given below.

RRT1: In this procedure, each interviewee in A with replacement simple random sample of size n is provided with one randomized response device. It consists of the statement “I belong to the sensitive group” with known probability P , exactly the same probability as used by Warner (1965) and the statement “Yes” with probability $(1 - P)$. The interviewee is instructed to use the device and report “Yes” or “No” for the random outcome of the sensitive statement according to his/her actual status. Otherwise, he is simply to report the “Yes” statement observed on the randomized response device. The whole procedure is completed by the respondent, unobserved by the interviewer. Then θ_1 , the probability of a “Yes” answer in the population, is

$$\theta_{s1} = P\pi + (1 - P) .$$

An unbiased estimator of π due to Singh (1993) is given by

$$\hat{\pi}_{s1} = \frac{\hat{\theta}_{s1} - (1 - P)}{P} ,$$

where $\hat{\theta}_{s1}$ is the proportion of “Yes” answers in the sample of size n .

The variance of the estimator $\hat{\pi}_{s1}$ is given by

$$V(\hat{\pi}_{s1}) = \frac{\pi(1 - \pi)}{n} + \frac{(1 - P)(1 - \pi)}{nP} . \quad (4)$$

RRT2: This procedure is exactly like RRT1 except for a change in probabilities on the randomized response device, i.e., the probabilities for the “sensitive” statement and “Yes” statement have been interchanged. The probability of a “Yes” response is then

$$\hat{\pi}_{s2} = \frac{\hat{\theta}_{s2} - P}{(1 - P)} ,$$

with variance is given by

$$V(\hat{\pi}_{s_2}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-\pi)}{n(1-P)} . \quad (5)$$

Huang (2004) showed that his procedure resulted in better performance as compared to the Warner (1965) and Chang and Huang (2001) procedures.

Huang Model

In this procedure, a simple random sample of size n is drawn with replacement from a finite population. The sampled individuals are required to reply to a direct query as to whether or not they belong to A. When answering “No”, the respondent is provided with a randomization device consisting of two statements:

- (i) “I am a member of A”
- (ii) “I am not a member of A”

with probabilities P and $(1 - P)$, respectively.

It is assumed that the respondents belonging to A give totally honest responses under the randomized response procedure, but with probability T following the usual direct response procedure. The probability of a “Yes” response in the direct response procedure is given by

$$\theta_1 = \pi T ,$$

and in the randomized response procedure by

$$\theta_2 = P\pi(1-T) + (1-P)(1-\pi) = (2P-1)\pi - P\pi T + (1-P) .$$

Huang (2004) suggested the following estimators of π and T respectively as

$$\hat{\pi}_H = \frac{[P\hat{\theta}_1 + \hat{\theta}_2 - (1-P)]}{(2P-1)}$$

and

$$\hat{T}_H = \frac{(2P-1)\hat{\theta}_1}{\left[P\hat{\theta}_1 + \hat{\theta}_2 - (1-P)\right]},$$

where $\hat{\theta}_j$, the observed proportion of “Yes” answers, is the binomial random variable with parameters n and θ_j , $j = 1, 2$. Huang (2004) obtained the variance of $\hat{\pi}_H$ as

$$V(\hat{\pi}_H) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)(1-\pi T)}{n(2P-1)^2} \quad (6)$$

and the MSE of the estimator $\hat{T}_H$, up to terms of order $O(n^{-1})$, as

$$\text{MSE}(\hat{T}_H) = \frac{T(1-T)}{n\pi} + \frac{P(1-P)T^2(1-\pi T)}{n(2P-1)^2\pi^2} \quad (7)$$

Proposed Procedures

HRRT1

In this procedure, a simple random sample of size n is drawn with replacement from a finite population. The sampled individuals are instructed to answer a direct query as to whether or not he/she belongs A. When answering “No”, the respondent is provided with a randomization device. It consists of the statement “I belong to the sensitive group” with known probability P , exactly the same probability as used by Warner (1965), and the statement “Yes” with probability $(1 - P)$ (Singh 1993, p. 68). The interviewee is instructed to use the device and report “Yes” or “No” for the random outcome of the sensitive statement according to his/her actual status. Otherwise, they are simply to report the “Yes” statement observed on the randomized response device. The whole procedure is completed by the respondent, unobserved by the interviewer. Then θ_{t1} , the probability of a “Yes” answer in the population, is

$$\theta_{t1} = \pi T \Rightarrow T = \frac{\theta_{t1}}{\pi}$$

And, adopting the randomized response procedure, the respondent gives totally honest responses under the randomized response procedure by

$$\theta_{t1} = P\pi(1-T) + (1-P) \Rightarrow \pi = \frac{[P\theta_{t1} + \theta_{t2} - (1-P)]}{P}.$$

Thus the proposed estimators of π and T are given by

$$\hat{\pi}_{a1} = \frac{[P\hat{\theta}_{t1} + \hat{\theta}_{t2} - (1-P)]}{P}$$

and

$$\hat{T}_1 = \frac{P\hat{\theta}_{t1}}{[P\hat{\theta}_{t1} + \hat{\theta}_{t2} - (1-P)]},$$

respectively, where $\hat{\theta}_{tj}$, the observed proportion of “Yes” answers, is the binomial random variable with parameters n and θ_{tj} , $j = 1, 2$. The principal properties of the estimator $\hat{\pi}_{a1}$ are outlined in the following theorem:

Theorem 1. The estimator $\hat{\pi}_{a1}$ is unbiased with the variance given by

$$V(\hat{\pi}_{a1}) = \frac{\pi(1-\pi)}{n} + \frac{(1-P)(1-\pi-\pi T)}{nP} \quad (8)$$

Proof. The unbiasedness follows from $E(\hat{\theta}_{tj}) = \theta_{tj}$, $j = 1, 2$. The variance of the estimator $\hat{\pi}_{a1}$ can be obtained as follows:

$$\begin{aligned}
V(\hat{\pi}_{a1}) &= \frac{1}{P^2} \left[P^2 V(\hat{\theta}_{t1}) + V(\hat{\theta}_{t2}) + 2P \text{cov}(\hat{\theta}_{t1}, \hat{\theta}_{t2}) \right] \\
&= \frac{1}{P^2} \left[\frac{P^2 \theta_{t1} (1 - \theta_{t1})}{n} + \frac{\theta_{t2} (1 - \theta_{t2})}{n} - \frac{2P \theta_{t1} \theta_{t2}}{n} \right] \\
&= \frac{1}{nP^2} \left[P^2 \theta_{t1} + \theta_{t2} - (P^2 \theta_{t1}^2 + \theta_{t2}^2 + 2P \theta_{t1} \theta_{t2}) \right] \\
&= \frac{1}{nP^2} \left[P^2 \theta_{t1} + \theta_{t2} - (P \theta_{t1} + \theta_{t2})^2 \right] \\
&= \frac{1}{nP^2} \left[P^2 \pi (1 - \pi) + P(1 - P)(1 - \pi - \pi T) \right] \\
&= \frac{1}{n} \left[\pi (1 - \pi) + \frac{(1 - P)(1 - \pi - \pi T)}{P} \right]
\end{aligned}$$

Hence the theorem.

An unbiased estimator of the variance $V(\hat{\pi}_{a1})$ can easily be obtained, which is given as follows:

Theorem 2. The unbiased estimator of $V(\hat{\pi}_{a1})$ is given by

$$V(\hat{\pi}_{a1}) = \frac{\hat{\pi}_{a1}(1 - \hat{\pi}_{a1})}{n - 1} + \frac{(1 - P)(1 - \hat{\pi}_{a1} - \hat{\theta}_{a1})}{(n - 1)P} \quad (9)$$

To form an idea about the sampling fluctuation of the direct estimator $\hat{\pi}_D$ from the sample itself, one has to develop an estimator of $\text{MSE}(\hat{\pi}_D)$. In fact, with the help of the proposed procedure, one can find an unbiased estimator of the MSE of $\hat{\pi}_{a1}$, which is presented in the following theorem:

Theorem 3. The unbiased estimator of the MSE of $\hat{\pi}_D$ is given by

$$\begin{aligned}\text{MSE}(\hat{\pi}_D) &= (\hat{\theta}_{t1} - \hat{\pi}_{a1})^2 + \frac{2\hat{\theta}_{t1}}{(n-1)} \left[1 - \hat{\pi}_{a1} - \frac{(1-P)}{p} \right] - \hat{V}(\hat{\pi}_{a1}) \\ &= (\hat{\theta}_{t1} - \hat{\pi}_{a1})^2 + \frac{2\hat{\theta}_{t1}}{(n-1)P} [P - P\hat{\theta}_{t1} - \hat{\theta}_{t2}] - \hat{V}(\hat{\pi}_{a1})\end{aligned}$$

Proof. The proof is straightforward and is therefore omitted.

To obtain the bias and MSE of the estimator $\hat{T}$, we write $d_1 = P\hat{\theta}_1$ and $d_2 = [P\hat{\theta}_1 + \hat{\theta}_2 - (1-P)]$, and it follows that $E(d_1) = P\pi T$ and $E(d_2) = \pi P$. The estimator $\hat{T}$ can then be represented as $\hat{T} = d_1/d_2$, and we have $T = E(d_1)/E(d_2)$. Furthermore, we define the following quantities:

$$e_1 = \frac{d_1 - E(d_1)}{E(d_1)}$$

and

$$e_2 = \frac{d_2 - E(d_2)}{E(d_2)},$$

assuming that $|e_1| < 1$ so that the function $(1 + e_2)^{-1}$ can be validly expanded as a power series. It can be easily proved that

$$\begin{aligned}E(e_1^2) &= \frac{\theta_{t1}(1-\theta_{t1})}{n\theta_{t1}^2} = \frac{(1-\theta_{t1})}{n\theta_{t1}} = \frac{\theta_{t1}(1-\theta_{t1})}{n\pi^2 T^2}, \\ E(e_2^2) &= \frac{P^2\theta_{t1}(1-\theta_{t1}) + \theta_{t2}(1-\theta_{t2}) - 2P\theta_{t1}\theta_{t2}}{nP^2\pi^2}, \\ E(e_1 e_2) &= \frac{P\theta_{t1}(1-\theta_{t1}) - \theta_{t1}\theta_{t2}}{nP\pi^2 T}\end{aligned}$$

and the estimation error of the estimator $\hat{T}_1$ can be expressed as

$$\begin{aligned}
\hat{T}_1 - T &= \frac{P\theta_{t1}(1+e_1) - \theta_{t1}\theta_{t2}}{P\pi(1+e_2)} - T \\
&= \frac{P\pi T}{P\pi}(1+e_1)(1+e_2)^{-1} - T \\
&= T(1+e_1)(1+e_2)^{-1} - T \\
\hat{T}_1 - T &= T(e_1 - e_2) + o_p(n^{-1/2})
\end{aligned}$$

We then state the following theorem:

Theorem 4. The MSE of the estimator $\hat{T}_1$, up to terms of order $o(n^{-1})$, is given by

$$\text{MSE}(\hat{T}_1) = \frac{T(1-T)}{n\pi} + \frac{(1-P)T^2(1+\pi-\pi T)}{nP\pi^2} \quad (10)$$

Proof. Consider

$$\begin{aligned}
\text{MSE}(\hat{T}_1) &= E(\hat{T}_1 - T)^2 \\
&\cong T^2 [E(e_1^2) - 2E(e_1 e_2) + E(e_2^2)] \\
&= \frac{T^2}{n} \left[\frac{\theta_{t1}(1-\theta_{t1})}{\pi^2 T^2} - \frac{2[P\theta_{t1}(1-\theta_{t1}) - \theta_{t1}\theta_{t2}]}{P\pi^2 T^2} \right. \\
&\quad \left. + \frac{P^2\theta_{t1}(1-\theta_{t1}) + \theta_{t2}(1-\theta_{t2}) - 2P\theta_{t1}\theta_{t2}}{P^2\pi^2} \right] \\
&= \frac{T^2}{n\pi^2} \left[\frac{\theta_{t1}(1-\theta_{t1})(1-T)^2}{T^2} + \frac{2\theta_{t1}\theta_{t2}(1-T)}{PT} + \frac{\theta_{t2}(1-\theta_{t2})}{P^2} \right] \\
&= \frac{T^2}{n\pi^2} \left[\frac{P^2\theta_{t1}(1-T)^2 + T^2\theta_{t2}}{P^2 T^2} - \frac{[\theta_{t1}P(1-T) - \theta_{t2}T]^2}{P^2 T^2} \right] \\
&= \frac{1}{n} \left[\frac{T(1-T)}{\pi} + \frac{T^2(1-P)}{\pi^2 P} [1 + \pi(1-T)] \right]
\end{aligned}$$

$$= \frac{T(1-T)}{n\pi} + \frac{(1-P)T^2(1+\pi-\pi T)}{nP\pi^2}$$

Hence the theorem.

HRRT2

In this proposed method, a simple random sample of size n is drawn with replacement from a finite population. The sampled individuals are required to reply a direct query as to whether or not they belong to A. When answering “No”, the respondent is provided with a randomization device consisting of the statement “I belong to the sensitive group” with known probability $(1 - P)$, exactly the same probability as used by Warner (1965), and the statement “Yes” with probability P (Singh 1993, p. 68). The interviewee is instructed to use the device and report “Yes” or “No” for the random outcome of the sensitive statement according to his/her actual status. Otherwise, they are simply to report the “Yes” statement observed on the randomized response device. The whole procedure is completed by the respondent, unobserved by the interviewer. The probability of a "Yes" answer in the population is then

$$\theta_{t1} = \pi T$$

and in the randomized response procedure is

$$\theta_{t3} = \pi(1-P)(1-T) + P .$$

The proposed estimators of π and T are given by

$$\hat{\pi}_{a2} = \frac{[(1-P)\hat{\theta}_{t1} + \hat{\theta}_{t3} - P]}{(1-P)}$$

and

$$\hat{T}_2 = \frac{(1-P)\hat{\theta}_{t1}}{[(1-P)\hat{\theta}_{t1} + \hat{\theta}_{t3} - P]} ,$$

where $\hat{\theta}_{t1}$ and $\hat{\theta}_{t3}$ are the observed proportion of “Yes” answers. The principal properties of the estimator $\hat{\pi}_{a2}$ are outlined in the following theorem:

Theorem 5. The estimator $\hat{\pi}_{a2}$ is unbiased with variance given by

$$V(\hat{\pi}_{a2}) = \frac{\pi(1-\pi)}{n} + \frac{P(1-\pi-\pi T)}{n(1-P)} \quad (11)$$

Proof. The unbiasedness follows from $E(\hat{\theta}_{t1}) = \theta_{t1}$ and $E(\hat{\theta}_{t3}) = \theta_{t3}$. The variance of the estimator $\hat{\pi}_{a2}$ can be obtained as follows:

$$\begin{aligned} V(\hat{\pi}_{a2}) &= \frac{1}{(1-P)^2} \left[(1-P)^2 V(\hat{\theta}_{t1}) + V(\hat{\theta}_{t3}) + 2(1-P) \text{cov}(\hat{\theta}_{t1}, \hat{\theta}_{t3}) \right] \\ &= \frac{1}{n(1-P)^2} \left[\left[(1-P)^2 \theta_{t1} + \theta_{t3} \right] - \left[(1-P) \theta_{t1} + \theta_{t3} \right]^2 \right] \\ &= \frac{1}{n(1-P)^2} \left[(1-P)^2 \pi(1-\pi) + P(1-P)(1-\pi-\pi T) \right] \\ &= \frac{1}{2} \left[\pi(1-\pi) + \frac{P(1-\pi-\pi T)}{(1-P)} \right] \end{aligned}$$

Hence the theorem.

An unbiased estimator of the variance $V(\hat{\pi}_{a2})$ can easily be obtained, which is given in the following theorem.

Theorem 6. The unbiased estimator of $V(\hat{\pi}_{a2})$ is given by

$$V(\hat{\pi}_{a2}) = \frac{\hat{\pi}_{a2}(1-\hat{\pi}_{a2})}{n-1} + \frac{(1-\hat{\pi}_{a2}-\hat{\theta}_{t1})P}{(n-1)(1-P)}.$$

To form an idea about the sampling fluctuation of the direct estimator $\hat{\pi}_{D2}$ from the sample itself, one has to develop an estimator of $\text{MSE}(\hat{\pi}_{a2})$. In fact, with

the help of the proposed procedure, one can find an unbiased estimator of the MSE of $\hat{\pi}_{D2}$, which is presented in the following theorem.

Theorem 7. The unbiased estimator of MSE of $\hat{\pi}_{D2}$ is given by

$$\begin{aligned} \text{MSE}(\hat{\pi}_D) &= (\hat{\theta}_{t1} - \hat{\pi}_{a2}) + \frac{2\hat{\theta}_{t1}}{(n-1)} \left[1 - \hat{\pi}_{a2} - \frac{P}{(1-P)} \right] - \hat{V}(\hat{\pi}_{a2}) \\ &= (\hat{\theta}_{t1} - \hat{\pi}_{a2})^2 + \frac{2\hat{\theta}_{t1}}{(n-1)(1-P)} \left[(1-P) - (1-P)\hat{\theta}_{t1} - \hat{\theta}_{t3} \right] - \hat{V}(\hat{\pi}_{a2}) \end{aligned}$$

Proof. The proof is straightforward and omitted.

Now to obtain the MSE of $\hat{T}_2$, we define $d_1^* = (1-P)\hat{\theta}_{t1}$ and $d_2^* = [(1-P)\hat{\theta}_{t1} + \hat{\theta}_{t3} - P]$. It follows that $E(d_1^*) = (1-P)\pi T$ and $E(d_2^*) = [(1-P)\theta_{t1} + \theta_{t3} - P]$. The estimator $\hat{T}_2$ can then be represented as $\hat{T}_2 = d_1^*/d_2^*$, and we therefore have $T = E(d_1^*)/E(d_2^*)$. We then define the following quantities:

$$e_1^* = \frac{d_1^* - E(d_1^*)}{E(d_1^*)}$$

and

$$e_2^* = \frac{d_2^* - E(d_2^*)}{E(d_2^*)},$$

assuming that $|e_1| < 1$ so that the function $(1 + e_2)^{-1}$ can be validly expanded as a power series. It can be verified that

$$\begin{aligned} E(e_1^{*2}) &= \frac{\theta_{t1}(1-\theta_{t1})}{n\theta_{t1}^2} = \frac{\theta_{t1}(1-\theta_{t1})}{n\pi^2 T^2} \\ E(e_2^{*2}) &= \frac{(1-P)^2 \theta_{t1}(1-\theta_{t1}) + \theta_{t3}(1-\theta_{t3}) - 2(1-P)\theta_{t1}\theta_{t3}}{n\pi^2 (1-P)^2} \\ E(e_1^* e_2^*) &= \frac{(1-P)\theta_{t1}(1-\theta_{t1}) - \theta_{t1}\theta_{t3}}{n\pi^2 T(1-P)} \end{aligned}$$

and the estimation error of the estimator $\hat{T}_2$ can be expressed as

$$\begin{aligned} \hat{T}_2 - T &= \frac{(1-P)\theta_{t1}(1-e_1^*) - \theta_{t1}\theta_{t3}}{(1-P)\pi(1+e_2^*)} - T \\ &= \frac{(1-P)\pi T}{(1-P)\pi} (1-e_1^*)(1+e_2^*)^{-1} - T \\ &= T(1-e_1^*)(1+e_2^*)^{-1} - T \\ \hat{T}_2 - T &= T(e_1^* - e_2^*) + o_p(n^{-1/2}) \end{aligned}$$

Theorem 8. The MSE of the estimator $\hat{T}_2$, up to terms of order $o(n^{-1})$, is given by

$$MSE(\hat{T}_2) = \frac{T(1-T)}{n\pi} + \frac{PT^2(1+\pi-\pi T)}{n(1-P)\pi^2} \quad (12)$$

Proof.

$$\begin{aligned} MSE(\hat{T}_2) &= E(\hat{T}_2 - T)^2 \\ &\cong T^2 E(e_1^* - e_2^*)^2 \\ &= T^2 [E(e_1^{*2}) - 2E(e_1^* e_2^*) + E(e_2^{*2})] \end{aligned}$$

$$\begin{aligned}
 &= \frac{T^2}{n} \left[\frac{\theta_{t1}(1-\theta_{t1})}{\pi^2 T^2} - \frac{2[(1-P)\theta_{t1}(1-\theta_{t1}) - \theta_{t1}\theta_{t3}]}{(1-P)\pi^2 T} \right. \\
 &\quad \left. + \frac{(1-P)^2 \theta_{t1}(1-\theta_{t1}) + \theta_{t3}(1-\theta_{t3}) - 2(1-P)\theta_{t1}\theta_{t3}}{(1-P)^2 \pi^2} \right] \\
 &= \frac{T^2}{n\pi^2} \left[\frac{\theta_{t1}(1-T)^2}{T^2} + \frac{\theta_{t3}}{(1-P)^2} - \left(\frac{\theta_{t1}(1-T)}{T} - \frac{\theta_{t3}}{(1-P)} \right)^2 \right] \\
 &= \frac{1}{n} \left[\frac{T(1-T)}{\pi} + \frac{T^2(1-T)}{\pi(1-P)} + \frac{T^2 P}{\pi^2(1-P)} \right] \\
 &= \frac{1}{n} \left[\frac{T(1-T)}{\pi} + \frac{T^2(1-T)}{\pi(1-P)} - \frac{T(1-T)}{\pi} + \frac{T(1-T)^2}{\pi} + \frac{PT^2}{\pi^2(1-P)} \right] \\
 &= \frac{1}{n} \left[\frac{T(1-T)}{\pi} - \frac{T^2(1-T)}{\pi} + \frac{(1-T)T^2}{\pi(1-P)} + \frac{PT^2}{\pi^2(1-P)} \right] \\
 &= \frac{1}{n} \left[\frac{T(1-T)}{\pi} + \frac{T^2(1-T)P}{\pi(1-P)} + \frac{PT^2}{\pi^2(1-P)} \right] \\
 &= \frac{1}{n} \left[\frac{T(1-T)}{\pi} + \frac{PT^2(1+\pi-\pi T)}{\pi^2(1-P)} \right]
 \end{aligned}$$

Hence the theorem.

Theoretical Comparisons

Comparisons of the proposed estimators $\hat{\pi}_{a1}$ and $\hat{\pi}_{a2}$ with Warner's estimator $\hat{\pi}_w$

From (3) and (8),

$$\begin{aligned}
 V(\hat{\pi}_w) - V(\hat{\pi}_{a1}) &= \frac{(1-P)}{n} \left[\frac{P}{(2P-1)^2} - \frac{1}{P} + \frac{\pi(1+t)}{P} \right] \\
 &= \frac{(1-P)}{n(2P-1)^2 P} \left[P^2 - (2P-1)^2 + \pi(1+T)(2P-1)^2 \right]
 \end{aligned} \tag{13}$$

which is positive if

$$\left[(3P-1)(1-P) + \pi(1+T)(2P-1)^2 \right] > 0 . \quad (14)$$

The condition (14) is always true as long as $P > 1/3$. Thus the proposed estimator is more efficient than the Warner's (1965) estimator $\hat{\pi}_w$ if $P > 1/3$.

It is further observed from (3) and (11) that $V(\hat{\pi}_w) - V(\hat{\pi}_{a_2}) > 0$ if

$$\frac{P}{n(2P-1)^2(1-P)} \left[(1-P)^2 - (2P-1)^2(1-\pi-\pi T) \right] > 0 ,$$

i.e. if $[P(2-3P) + (2P-1)^2\pi(1+T)] > 0$, which is always true if $P < 2/3$. Thus the proposed estimator $\hat{\pi}_{a_2}$ is better than Warner's (1965) estimator as long as $P < 2/3$.

Comparisons of the proposed estimators $(\hat{\pi}_{a_1}, \hat{T}_1)$ with Huang's estimator $(\hat{\pi}_H, \hat{T}_H)$

From (7) and (8),

$$\begin{aligned} V(\hat{\pi}_H) - V(\hat{\pi}_{a_1}) &= \frac{(1-P)}{n} \left[\frac{(1-\pi T)}{(2P-1)^2 P} \left[P^2 - (2P-1)^2 \right] + \frac{\pi}{P} \right] \\ &= \frac{(1-P)}{nP(2P-1)^2} \left[(1-\pi T)(3P-1)(1-P) + \pi(2P-1)^2 \right] \quad (15) \\ &= \frac{(1-P)}{nP(2P-1)^2} \left[\begin{aligned} &(3P-1)(1-P) \\ &- \pi \left[T(3P-1)(1-P) - (2P-1)^2 \right] \end{aligned} \right] \end{aligned}$$

which is positive if

$$\left[(3P-1)(1-P) - \pi \left(T(3P-1)(1-P) - (2P-1)^2 \right) \right] > 0 ,$$

i.e. if either

$$\begin{aligned} \pi &< \frac{(3P-1)(1-P)}{\left[(3P-1)(1-P)T - (2P-1)^2\right]}, \quad P > \frac{1}{3} \quad \text{and} \\ T &> \frac{(2P-1)^2}{(3P-1)(1-P)} \end{aligned} \quad (16)$$

or

$$\begin{aligned} \pi &> \frac{(3P-1)(1-P)}{\left[(3P-1)(1-P)T - (2P-1)^2\right]}, \quad P < \frac{1}{3} \quad \text{and} \\ T &> \frac{(2P-1)^2}{(3P-1)(1-P)} \end{aligned} \quad (17)$$

Thus the proposed estimator $\hat{\pi}_{a1}$ is more efficient than Huang's (2004) estimator $\hat{\pi}_H$ as long as either inequality (16) or (17) is satisfied.

We note from (15) that the difference $V(\hat{\pi}_H) - V(\hat{\pi}_{a1})$ is always positive if

$$\begin{aligned} P^2 - (2P-1)^2 &> 0 \\ (3P-1)(1-P) &> 0 \end{aligned}$$

i.e. if

$$P > \frac{1}{3}, \quad (18)$$

which is a sufficient condition for the proposed estimator $\hat{\pi}_{a1}$ to be more efficient than Huang's (2004) estimator $\hat{\pi}_H$.

From (7) and (11) we have

$$\begin{aligned}
\text{MSE}(\hat{\pi}_H) - \text{MSE}(\hat{T}_1) &= \frac{P(1-P)T^2(1-\pi T)}{n(2P-1)^2\pi^2} - \frac{(1-P)(1+\pi-\pi T)T^2}{nP\pi^2} \\
&= \frac{(1-P)T^2}{n\pi^2} \left[\frac{P(1-\pi T)}{(2P-1)^2} - \frac{(1+\pi-\pi T)}{P} \right] \\
&= \frac{(1-P)T^2}{n(2P-1)^2\pi^2} \left[P^2(1-\pi T) - (2P-1)^2(1+\pi-\pi T) \right]
\end{aligned}$$

which is positive if

$$\begin{aligned}
&\left[P^2(1-\pi T) - (2P-1)^2(1+\pi-\pi T) \right] > 0 \\
&\left[P^2 - (2P-1)^2 - \pi \left[TP^2 + (1-T)(2P-1)^2 \right] \right] > 0 \\
&\left[P^2 - (2P-1)^2 - \pi \left[T(P^2 - (2P-1)^2) + (2P-1)^2 \right] \right] > 0 \\
&\left[(3P-1)(1-P) - \pi \left[T(3P-1)(1-P) + (2P-1)^2 \right] \right] > 0
\end{aligned}$$

i.e. if either

$$\pi < \frac{(3P-1)(1-P)}{\left[(2P-1)^2 + (3P-1)(1-P)T \right]}, \quad P > \frac{1}{3} \quad (19)$$

or

$$\pi > \frac{(3P-1)(1-P)}{\left[(2P-1)^2 + (3P-1)(1-P)T \right]}, \quad P < \frac{1}{3}. \quad (20)$$

It follows that the proposed estimator $\hat{T}_1$ is better than Huang's (2004) estimator $\hat{T}_H$ if either (19) or (20) holds.

Comparisons of the proposed estimators $\hat{\pi}_{a2}$ and $\hat{T}_2$ with Huang's estimator $(\hat{\pi}_H, \hat{T}_H)$

From (6) and (11),

$$\begin{aligned} V(\hat{\pi}_H) - V(\hat{\pi}_{a2}) &= \frac{P(1-P)}{n(2P-1)^2} - \frac{P(1-\pi-\pi T)}{n(1-P)} \\ &= \frac{P}{n(1-P)(2P-1)^2} \left[\pi(TP(3P-2) + (2P-1)^2) - P(3P-2) \right] \end{aligned}$$

Which is positive if either

$$\pi > \frac{P(3P-2)}{\left[(2P-1)^2 + P(3P-2)T \right]}, \quad P > \frac{2}{3} \quad (21)$$

or

$$\pi < \frac{P(3P-2)}{\left[(2P-1)^2 + P(3P-2)T \right]}, \quad P < \frac{2}{3}. \quad (22)$$

Thus the proposed estimator $\hat{\pi}_{a2}$ is better than Huang's (2004) estimator $\hat{T}_H$ if either (21) or (22) holds.

From (18) it is further observed that the different $V(\hat{\pi}_H) - V(\hat{\pi}_{a2})$ is always positive if $2 - 3P > 0$, i.e. if

$$P < \frac{2}{3}. \quad (23)$$

The condition (23) is sufficient for the proposed estimator $\hat{\pi}_{a2}$ to be better than Huang's (2004) estimator $\hat{T}_H$.

From (7) and (12),

$$\begin{aligned} \text{MSE}(\hat{T}_H) - V(\hat{T}_2) &= \frac{PT^2}{n\pi^2} \left[\frac{(1-P)(1-\pi T)}{(2P-1)^2} - \frac{(1+\pi-\pi T)}{(1-P)} \right] \\ &= \frac{PT^2}{n\pi^2(1-P)(2P-1)^2} \left[P(2-3P) - \pi \left[P(2-3P)T + (2P-1)^2 \right] \right] \end{aligned}$$

which is positive if either

$$\pi < \frac{P(2-3P)}{\left[P(2-3P)T + P(2P-1)^2 \right]}, \quad P < \frac{2}{3} \quad (24)$$

or

$$\pi > \frac{P(2-3P)}{\left[P(2-3P)T + P(2P-1)^2 \right]}, \quad P > \frac{2}{3}. \quad (25)$$

It follows that the proposed estimator $\hat{T}_2$ is more efficient than Huang's (2004) estimator $\hat{T}_H$ if either (24) or (25) is satisfied.

Comparisons of the proposed estimators $(\hat{\pi}_{a2}, \hat{T}_1)$ with the proposed estimators $(\hat{\pi}_{a2}, \hat{T}_2)$

From (7) and (11),

$$V(\hat{\pi}_{a2}) - V(\hat{\pi}_{a1}) = \frac{(1-\pi-\pi T)(2P-1)}{nP(1-P)},$$

which is positive if either

$$\pi < \frac{1}{1+T}, \quad P > \frac{1}{2} \quad (26)$$

or

$$\pi < \frac{1}{1+T}, \quad P < \frac{1}{2} \quad (27)$$

Thus the proposed estimator $\hat{\pi}_{a1}$ is more efficient than the proposed estimator $\hat{\pi}_{a2}$ if either (26) or (27) is satisfied.

Further, the difference $V(\hat{\pi}_{a2}) - V(\hat{\pi}_{a1}) < 0$ if

$$\frac{(1-\pi-\pi T)(2P-1)}{nP(1-P)} < 0$$

i.e. if $[1 - \pi(1 + T)](2P - 1) < 0$,

$$\pi > \frac{1}{1+T}, \quad P > \frac{1}{2} \quad (28)$$

or

$$\pi < \frac{1}{1+T}, \quad P < \frac{1}{2} \quad (29)$$

It follows from the above that the proposed estimator $\hat{\pi}_{a2}$ is more efficient than the suggested estimator $\hat{\pi}_{a1}$ if either (28) or (29) is satisfied.

From (10) and (12) we have

$$\text{MSE}(\hat{T}_2) - \text{MSE}(\hat{T}_1) = \frac{T^2}{n\pi^2} (1 + \pi - \pi T) \left[\frac{P}{1-P} - \frac{(1-P)}{P} \right],$$

which is positive if $P^2 - (1 - P)^2 > 0$, i.e. if

$$P > \frac{1}{2}. \quad (30)$$

Thus if $P > 1/2$ holds, the proposed estimator $\hat{T}_1$ is better than the estimator $\hat{T}_2$. On the other hand, for $P < 1/2$, the proposed estimator $\hat{T}_2$ would be better than $\hat{T}_1$.

Comparisons of the proposed estimators $(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ with direct estimator $\hat{\pi}_D$.

From (2) and (7) we have

$$V(\hat{\pi}_D) - V(\hat{\pi}_{a1}) = \frac{1}{nP} \left[\begin{array}{l} nP\pi^2(1-T)^2 + \pi T(1-\pi T)P \\ - P\pi(1-\pi) - (1-P)(1-\pi-\pi T) \end{array} \right]$$

which is greater than zero if

$$\left[nP\pi^2(1-T)^2 + \pi T(1-\pi T)P - P\pi(1-\pi) - (1-P)(1-\pi-\pi T) \right] > 0$$

i.e. if

$$n > \frac{1}{P\pi^2(1-T)^2} \left[P\pi(1-T)(1-\pi-\pi T) + (1-P)(1-\pi-\pi T) \right]$$

i.e. if

$$n > \frac{(1-\pi-\pi T)}{P\pi^2(1-T)^2} \left[P\pi(1-T) + (1-P) \right], \quad \pi < \frac{1}{(1+T)}. \quad (31)$$

Thus the proposed estimator $\hat{\pi}_{a1}$ is more efficient than the direct estimator $\hat{\pi}_D$ if the inequality (31) holds.

From (2) and (11) we have

$$\text{MSE}(\hat{\pi}_D) - \text{MSE}(\hat{\pi}_{a2}) = \frac{1}{n(1-P)} \left[\begin{array}{l} n(1-P)\pi^2(1-T)^2 + \pi T(1-\pi T)(1-P) \\ - (1-P)\pi(1-\pi) - P(1-\pi-\pi T) \end{array} \right],$$

which is greater than zero if

$$n(1-P)\pi^2(1-T)^2 + \pi T(1-\pi T)(1-P) - (1-P)\pi(1-\pi) - P(1-\pi-\pi T) > 0$$

i.e. if

$$n > \frac{1}{(1-P)\pi^2(1-T)^2} \left[(1-P)\pi(1-\pi) + P(1-\pi-\pi T) - (1-P)\pi T(1-\pi T) \right]$$

$$n > \frac{1}{(1-P)\pi^2(1-T)^2} \left[(1-P)\pi(1-T)(1-\pi-\pi T) + P(1-\pi-\pi T) \right]$$

$$n > \frac{(1-\pi-\pi T)}{(1-P)\pi^2(1-T)^2} \left[(1-P)\pi(1-T) + P \right]$$

i.e. if

$$\pi < \frac{1}{(1+T)} . \quad (32)$$

It follows that the proposed estimator $\hat{\pi}_{a2}$ is more efficient than the direct estimator $\hat{\pi}_D$ if the condition (32) holds.

Numerical Illustration

This illustration is provided to give a tangible idea about the magnitude of the relative efficiency of the suggested procedures with respect to the Huang (2004) and direct estimator procedures. The percent relative efficiency (PRE) of the proposed estimators ($\hat{\pi}_{a1}, \hat{\pi}_{a2}$) in relation to the Huang (2004) estimator $\hat{\pi}_H$ are given by

$$\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_H) = \frac{P \left[(2P-1)^2 \pi(1-\pi) + P(1-P)(1-\pi T) \right]}{(2P-1)^2 \left[P\pi(1-\pi) + (1-P)(1-\pi-\pi T) \right]} \times 100 \quad (33)$$

and

$$\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H) = \frac{(1-P) \left[(2P-1)^2 \pi(1-\pi) + P(1-P)(1-\pi T) \right]}{(2P-1)^2 \left[(1-P)\pi(1-\pi) + P(1-\pi-\pi T) \right]} \times 100 , \quad (34)$$

respectively. The PRE of the proposed estimators $(\hat{T}_1, \hat{T}_2)$ with respect to the Huang (2004) estimator $\hat{T}_H$ are given by

$$\text{PRE}(\hat{T}_1, \hat{T}_H) = \frac{P[(1-T)(2P-1)^2 \pi + P(1-P)T(1-\pi T)]}{(2P-1)^2 [(1-T)P\pi + (1-P)T(1+\pi - \pi T)]} \times 100 \quad (35)$$

and

$$\text{PRE}(\hat{T}_2, \hat{T}_H) = \frac{(1-P)[(1-T)(2P-1)^2 \pi + TP(1-P)(1-\pi T)]}{(2P-1)^2 [(1-T)(1-P)\pi + PT(1+\pi - \pi T)]} \times 100, \quad (36)$$

respectively. The expression for PRE of the proposed estimators $(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ in relation to the direct estimator $\hat{\pi}_D$ are given by

$$\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D) = \frac{P[\pi T(1-\pi T) + n\pi^2(1-T)^2]}{P\pi(1-\pi) + (1-P)(1-\pi - \pi T)} \times 100 \quad (37)$$

and

$$\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D) = \frac{(1-P)[\pi T(1-\pi T) + n\pi^2(1-T)^2]}{(1-P)\pi(1-\pi) + P(1-\pi - \pi T)} \times 100, \quad (38)$$

respectively. For comparing the two proposed procedures, we give the PRE of $(\hat{\pi}_{a1}, \hat{T}_1)$ with respect to $(\hat{\pi}_{a2}, \hat{T}_2)$ as

$$\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2}) = \frac{P[(1-P)\pi(1-\pi) + P(1-\pi - \pi T)]}{[P\pi(1-\pi) + (1-P)(1-\pi - \pi T)](1-P)} \times 100 \quad (39)$$

and

$$\text{PRE}(\hat{T}_1, \hat{T}_2) = \frac{P[(1-P)\pi T(1-T) + PT^2(1+\pi-\pi T)]}{[P\pi T(1-T) + (1-P)T^2(1+\pi-\pi T)](1-P)} \times 100, \quad (40)$$

respectively. We have further obtained the expressions for PRE of the proposed estimators $(\hat{\pi}_{a2}, \hat{T}_2)$ with respect to $(\hat{\pi}_{a1}, \hat{T}_1)$, given by

$$\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_{a1}) = \frac{(1-P)[P\pi(1-\pi) + (1-P)(1-\pi-\pi T)]}{P[(1-P)\pi(1-\pi) + P(1-\pi-\pi T)]} \times 100 \quad (41)$$

and

$$\text{PRE}(\hat{T}_2, \hat{T}_1) = \frac{(1-P)[P\pi T(1-T) + (1-P)T^2(1+\pi-\pi T)]}{P[(1-P)\pi T(1-T) + PT^2(1+\pi-\pi T)]} \times 100, \quad (42)$$

Respectively.

Using the formulae (33)-(42), we have computed

- (i) The $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_1, \hat{T}_H)$ for the values of $P = 0.6, 0.7, 0.8$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; and $\pi = 0.1 (0.1) 0.9$. The results are displayed in Tables 1 and 3.
- (ii) The $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_2, \hat{T}_H)$ for the values of $P = 0.20, 0.30, 0.40$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; and $\pi = 0.1 (0.1) 0.9$. The results are shown in Tables 2 and 4.
- (iii) The $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$ for the values of $P = 0.60, 0.70, 0.80$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; $\pi = 0.1 (0.1) 0.9$; and $n = 1000$. The findings are displayed in Table 5.
- (iv) The $\text{PRE}(\hat{\pi}_{a2}, \hat{T}_D)$ for the values of $P = 0.2, 0.30, 0.40$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; $\pi = 0.1 (0.1) 0.9$; and $n = 1000$. The findings are displayed in Table 6.
- (v) The $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ and $\text{PRE}(\hat{T}_1, \hat{T}_2)$ for the values of $P = 0.6, 0.7, 0.8$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; and $\pi = 0.1 (0.1) 0.9$. The findings are displayed in Tables 7 and 8.

- (vi) The $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_{a1})$ and $\text{PRE}(\hat{T}_2, \hat{T}_1)$ for the values of $P = 0.20, 0.30, 0.40$; $T = 0.10, 0.15, 0.30, 0.45, 0.60$; and $\pi = 0.1 (0.1) 0.9$. The results are displayed in Tables 9 and 10.

Tables 1 and 3 show that the values of $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_1, \hat{T}_H)$ are larger than 100, showing that $\hat{\pi}_H$ is more efficient than $\hat{\pi}_{a1}$ and that $\hat{T}_1$ is also superior to $\hat{T}_H$. It is observed that, for fixed values of (P, T) , the $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_H) \left[\text{PRE}(\hat{T}_1, \hat{T}_H) \right]$ decreases (increases) as π increases (decreases) slowly (rapidly). For fixed values of (T, π) , both $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_1, \hat{T}_H)$ decrease in a speedy manner as P increases. Thus a larger gain in efficiency by using $\hat{\pi}_{a1} (\hat{T}_1)$ over $\hat{\pi}_H (\hat{T}_H)$ is expected when P is close to 0.5.

Tables 2 and 4 exhibit that

- For fixed values of (P, T) , the $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H) \left[\text{PRE}(\hat{T}_2, \hat{T}_H) \right]$ increases (decreases) as π increases (decreases)
- $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H) \left[\text{PRE}(\hat{T}_2, \hat{T}_H) \right]$ decreases (increases) slowly (rapidly) as T increases (decreases)
- Both $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_2, \hat{T}_H)$ increase in a speedy way as P increases

Hence higher gain in efficiency by using $\hat{\pi}_{a2} (\hat{T}_2)$ is observed when (P, π) are closer to 0.5. It is further observed from Tables 2 and 4 that the values of $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_H)$ and $\text{PRE}(\hat{T}_2, \hat{T}_H)$ are greater than 100, from which it follows that the envisaged estimators $(\hat{\pi}_{a2}, \hat{T}_2)$ are more efficient than Huang's (2004) estimators $(\hat{\pi}_H, \hat{T}_H)$.

It is observed from Table 5 that

- For fixed (P, T) , the $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$ increases as π increases

NEW PROCEDURES OF ESTIMATING PROPORTION USING RRT

- For fixed (P, π) , the $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$ decreases as T increases
- For fixed (T, π) , the $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$ increases as P increases

There is substantial gain in efficiency through use of the proposed estimator $\hat{\pi}_{a1}$ over direct estimator $\hat{\pi}_D$ for all values of (P, π, T) considered here.

It is observed from Table 6 that

- For fixed (P, T) , the $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D)$ increases as π increases
- For fixed (P, π) , the $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D)$ decreases as T increases
- For fixed (T, π) , the $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D)$ increases as P increases

There is substantial gain in efficiency through use of the proposed estimator $\hat{\pi}_{a2}$ over direct estimator $\hat{\pi}_D$ for all values of (P, π, T) considered here.

Tables 5 and 6 clearly demonstrate the superiority of the proposed estimators $\hat{\pi}_{a1}$ and $\hat{\pi}_{a2}$ over the usual direct estimator $\hat{\pi}_D$ as the values of $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$ and $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D)$ are larger than 100 for all values of (P, π, T) considered.

Tables 7 and 8 demonstrate that the values of $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ and $\text{PRE}(\hat{T}_1, \hat{T}_2)$ are greater than 100 for $0.10 \leq T \leq 0.60$, $0.10 \leq T \leq 0.50$, and $P > 1/2$. Both $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ and $\text{PRE}(\hat{T}_1, \hat{T}_2)$ increase in a speedy way as P increases. Hence higher gain in efficiency by using $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2})$ and $\text{PRE}(\hat{T}_1, \hat{T}_2)$ is observed when (P, π) are closer to 0.5.

It is observed from Tables 9 and 10 that the values of $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_{a1})$ and $\text{PRE}(\hat{T}_2, \hat{T}_1)$ are greater than 100 for $0.10 \leq T \leq 0.60$, $0.10 \leq T \leq 0.50$, and $P < 1/2$. Thus the proposed procedures $(\hat{\pi}_{a2}, \hat{T}_2)$ are more efficient than the estimators $(\hat{\pi}_{a1}, \hat{T}_1)$. Higher gains in efficiencies are observed for lower values of P (i.e. for the values of P close to zero).

Finally we conclude that the proposed procedures are superior to the Huang (2004) procedure and hence the Chang and Huang (2001) procedure, and to the usual direct procedure.

Conclusion

Randomized response procedures are attractive mechanisms for counteracting fears in response and providing with valid statistical inferences concerning a population. The proposed randomized response procedure allows us to estimate the population proportion π unbiasedly and to get an admissible estimator for T , which is an unattainable feature for most of the competing methods. It has been shown theoretically and empirically that the proposed procedures are better than the Warner (1965), Chang and Huang (2001), and Huang (2004) procedures. The unbiased estimators of the MSE are provided for the direct response survey based on the proposed RR techniques. The suggested procedure is therefore recommended for application in survey sampling practice.

References

- Chang, H. J. & Huang K. C. (2001). Estimation of proportion and sensitivity of a qualitative character. *Metrika*, 53(3), 269-280. doi: [10.1007/s001840100109](https://doi.org/10.1007/s001840100109)
- Chang, H. J., Wang, C. L., & Huang, K. C. (2004a). On estimating the proportion of a qualitative sensitive character using randomized response sampling. *Quality and Quantity*, 38(5), 675-680. doi: [10.1007/s11135-005-8105-4](https://doi.org/10.1007/s11135-005-8105-4)
- Chang, H. J., Wang, C. L., & Huang, K. C. (2004b). Using randomized response to estimate the proportion and truthful reporting probability in a dichotomous finite population. *Journal of Applied Statistics*, 31(5), 565-573. doi: [10.1080/02664760410001681819](https://doi.org/10.1080/02664760410001681819)
- Chua, T. C. & Tsui, A. K. (2000). Procuring honest responses indirectly. *Journal of Statistics Planning and Inference*, 90(1), 107-116. doi: [10.1016/S0378-3758\(00\)00109-9](https://doi.org/10.1016/S0378-3758(00)00109-9)
- Fox, J. A. & Tracy, P. E. (1986). *Randomized response: A method of sensitive surveys*. Newbury Park, CA: SEGE Publications.
- Huang, K. C. (2004). Survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *Statistica Neerlandica*, 58(1), 75-82. doi: [10.1046/j.0039-0402.2003.00113.x](https://doi.org/10.1046/j.0039-0402.2003.00113.x)
- Mahmood, M., Singh, S. & Horn, S. (1998). On the confidentiality guaranteed under randomized response sampling: A comparison with several new techniques. *Biomedical Journal*, 40(2), 237-242. doi: [10.1002/\(SICI\)1521-4036\(199806\)40:2<237::AID-BIMJ237>3.0.CO;2-N](https://doi.org/10.1002/(SICI)1521-4036(199806)40:2<237::AID-BIMJ237>3.0.CO;2-N)

NEW PROCEDURES OF ESTIMATING PROPORTION USING RRT

- Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society. Series B*, 56(1), 93-95.
- Mangat, N. S. & Singh, R. (1990). An alternative randomized procedure. *Biometrika*, 77(2), 439-442. doi: [10.1093/biomet/77.2.439](https://doi.org/10.1093/biomet/77.2.439)
- Sing, S., Singh, R., and Mangat, N. S. (2000). Some alternative strategies to Moor's model in randomized response model. *Journal of Statistical Planning and Inference*, 83(1), 243-255.
- Singh, H. P. & Tarray, T. A. (2012). A stratified unknown repeated trials in randomized response sampling. *Communications for Statistical Applications and Methods*, 19(6), 751-759. doi: [10.5351/CKSS.2012.19.6.751](https://doi.org/10.5351/CKSS.2012.19.6.751)
- Singh, H. P. & Tarray, T. A. (2013a). A modified survey technique for estimating the proportion and sensitivity in a dichotomous finite population. *International Journal of Advanced Scientific and Technical Research*, 3(6), 459-472.
- Singh, H. P. & Tarray, T. A. (2013b). An alternative to Kim and Warde's mixed randomized response model. *Statistics and Operations Research Transactions*, 37(2), 189-210.
- Singh, H. P. & Tarray, T. A. (2013c). An alternative to Kim and Warde's mixed randomized response technique. *Statistica*, 73(3), 379-402.
- Singh, H. P. & Tarray, T. A. (2014a). An alternative to stratified Kim and Warde's randomized response model using optimal (Neyman) allocation. *Model Assisted Statistics and Applications*, 9(1), 37-62.
- Singh, H. P. & Tarray, T. A. (2014b). An improved mixed randomized response model. *Model Assisted Statistics and Applications*, 9(1), 73-87.
- Singh, H. P. & Tarray, T. A. (2014c). A dexterous randomized response model for estimating a rare sensitive attribute using Poisson distribution. *Statistics & Probability Letters*, 90, 42-45. doi: [10.1016/j.spl.2014.03.019](https://doi.org/10.1016/j.spl.2014.03.019)
- Singh, H. P. & Tarray, T. A. (2014d). A modified mixed randomized response model. *Statistics in Transition new series*, 15(1), 67-82.
- Singh, H. P. & Tarray, T. A. (2014e). An improvement over Kim and Elam stratified unrelated question randomized response model using Neyman allocation. *Sankhya B*, 77(1). doi: [10.1007/s13571-014-0088-5](https://doi.org/10.1007/s13571-014-0088-5)
- Singh, R. & Mangat, N. S. (1996). *Elements of survey sampling* (Vol. 15). Dordrecht, Netherlands: Kluwer Academic Publishers.
- Singh, S. (1993). An alternative to Warner's randomized response technique. *Statistica*, 53(1), 67-71.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309), 63-69. doi: [10.1080/01621459.1965.10480775](https://doi.org/10.1080/01621459.1965.10480775)

Appendix

Table 1. The present relative efficiency of the proposed estimator $\hat{\pi}_{a1}$ with respect to Huang (2004) estimator $\hat{\pi}_H$, i.e. $PRE(\hat{\pi}_{a1}, \hat{\pi}_H)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.6	0.10	882.44	888.24	918.27	978.26	1081.82
0.6	0.15	882.35	888.12	918.56	980.00	1087.50
0.6	0.30	882.09	887.76	919.46	985.71	1106.90
0.6	0.45	881.82	887.37	920.45	992.31	1130.77
0.6	0.60	881.54	886.96	921.56	1000.00	1160.87
0.7	0.10	294.72	292.59	298.33	312.50	338.00
0.7	0.15	294.66	292.47	298.23	312.61	338.79
0.7	0.30	294.50	292.10	297.90	312.98	341.41
0.7	0.45	294.34	291.71	297.54	313.39	344.48
0.7	0.60	294.17	291.30	297.15	313.86	348.14
0.8	0.10	169.60	167.76	169.83	175.44	185.44
0.8	0.15	169.57	167.69	169.75	175.41	185.58
0.8	0.30	169.47	167.47	169.50	175.31	186.01
0.8	0.45	169.36	167.24	169.24	175.20	186.49
0.8	0.60	169.26	167.00	168.95	175.08	187.04

Table 2. The present relative efficiency of the proposed estimator $\hat{\pi}_{a2}$ with respect to Huang (2004) estimator $\hat{\pi}_H$, i.e. $PRE(\hat{\pi}_{a2}, \hat{\pi}_H)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.2	0.10	169.60	167.76	169.83	175.44	185.44
0.2	0.15	169.57	167.69	169.75	175.41	185.58
0.2	0.30	169.47	167.47	169.50	175.31	186.01
0.2	0.45	169.36	167.24	169.24	175.20	186.49
0.2	0.60	169.26	167.00	168.95	175.08	187.04
0.3	0.10	294.72	292.59	298.33	312.50	338.00
0.3	0.15	294.66	292.47	298.23	312.61	338.79
0.3	0.30	294.50	292.10	297.90	312.98	341.41
0.3	0.45	294.34	291.71	297.54	313.39	344.48
0.3	0.60	294.17	291.30	297.15	313.86	348.14
0.4	0.10	882.44	888.24	918.27	978.26	1081.82
0.4	0.15	882.35	888.12	918.56	980.00	1087.50
0.4	0.30	882.09	887.76	919.46	985.71	1106.90
0.4	0.45	881.82	887.37	920.45	992.31	1130.77
0.4	0.60	881.54	886.96	921.56	1000.00	1160.87

Table 3. The present relative efficiency of the proposed estimator $\hat{t}_1$ with respect to Huang (2004) estimator $\hat{t}_H$, i.e. $PRE(\hat{t}_1, \hat{t}_H)$.

P	T	π				
		0.1	0.2	0.3	0.4	0.5
0.6	0.10	420.49	296.91	240.23	207.69	186.59
0.6	0.15	502.07	363.41	292.90	250.21	221.59
0.6	0.30	639.44	497.83	408.85	347.76	303.23
0.6	0.45	708.88	579.46	486.01	415.36	360.08
0.6	0.60	750.88	634.38	540.85	464.10	400.00
0.7	0.10	160.87	133.85	122.47	116.19	112.21
0.7	0.15	180.24	147.21	131.98	123.21	117.52
0.7	0.30	217.73	178.01	155.46	140.92	130.77
0.7	0.45	239.51	199.79	173.45	154.70	140.67
0.7	0.60	253.80	216.09	187.68	165.51	147.72
0.8	0.10	114.29	106.71	103.77	102.20	101.23
0.8	0.15	119.87	109.72	105.50	103.18	101.72
0.8	0.30	132.67	117.66	110.17	105.67	102.68
0.8	0.45	141.63	124.32	114.17	107.49	102.76
0.8	0.60	148.30	130.03	117.59	108.58	101.75

Table 4. The present relative efficiency of the proposed estimator $\hat{t}_2$ with respect to Huang (2004) estimator $\hat{t}_H$, i.e. $PRE(\hat{t}_2, \hat{t}_H)$.

P	T	π				
		0.1	0.2	0.3	0.4	0.5
0.2	0.10	114.29	106.71	103.77	102.20	101.23
0.2	0.15	119.87	109.72	105.50	103.18	101.72
0.2	0.30	132.67	117.66	110.17	105.67	102.68
0.2	0.45	141.63	124.32	114.17	107.49	102.76
0.2	0.60	148.30	130.03	117.59	108.58	101.75
0.3	0.10	160.87	133.85	122.47	116.19	112.21
0.3	0.15	180.24	147.21	131.98	123.21	117.52
0.3	0.30	217.73	178.01	155.46	140.92	130.77
0.3	0.45	239.51	199.79	173.45	154.70	140.67
0.3	0.60	253.80	216.09	187.68	165.51	147.72
0.4	0.10	420.49	296.91	240.23	207.69	186.59
0.4	0.15	502.07	363.41	292.90	250.21	221.59
0.4	0.30	639.44	497.83	408.85	347.76	303.23
0.4	0.45	708.88	579.46	486.01	415.36	360.08
0.4	0.60	750.88	634.38	540.85	464.10	400.00

NEW PROCEDURES OF ESTIMATING PROPORTION USING RRT

Table 5. The present relative efficiency of the proposed estimator $\hat{\pi}_{a1}$ with respect to the direct estimator $\hat{\pi}_D$, i.e. $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_D)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.6	0.10	1186.81	4767.59	11105.95	21136.70	36826.82
0.6	0.15	1064.67	4296.40	10062.06	19276.07	33880.20
0.6	0.30	735.69	3008.63	7164.63	14018.86	25371.21
0.6	0.45	464.84	1923.46	4660.53	9336.08	17492.16
0.6	0.60	254.83	1060.70	2613.34	5371.33	10489.57
0.7	0.10	1720.28	6558.88	14669.65	27008.00	45736.53
0.7	0.15	1542.72	5903.90	13259.85	24533.18	41813.57
0.7	0.30	1064.93	4119.60	9371.92	17613.44	30656.88
0.7	0.45	672.17	2623.79	6047.15	11558.95	20605.66
0.7	0.60	368.09	1441.11	3360.83	6539.01	11977.45
0.8	0.10	2595.17	9132.28	19318.97	34115.37	55875.17
0.8	0.15	2326.03	8206.84	17409.49	30841.71	50721.23
0.8	0.30	1602.96	5697.51	12188.11	21807.11	36334.07
0.8	0.45	1010.03	3609.45	7784.14	14071.77	23780.20
0.8	0.60	552.13	1971.39	4278.71	7812.85	13403.33

Table 6. The present relative efficiency of the proposed estimator $\hat{\pi}_{a2}$ with respect to the direct estimator $\hat{\pi}_D$, i.e. $\text{PRE}(\hat{\pi}_{a2}, \hat{\pi}_D)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.2	0.10	2595.17	9132.28	19318.97	34115.37	55875.17
0.2	0.15	2326.03	8206.84	17409.49	30841.71	50721.23
0.2	0.30	1602.96	5697.51	12188.11	21807.11	36334.07
0.2	0.45	1010.03	3609.45	7784.14	14071.77	23780.20
0.2	0.60	552.13	1971.39	4278.71	7812.85	13403.33
0.3	0.10	1720.28	6558.88	14669.65	27008.00	45736.53
0.3	0.15	1542.72	5903.90	13259.85	24533.18	41813.57
0.3	0.30	1064.93	4119.60	9371.92	17613.44	30656.88
0.3	0.45	672.17	2623.79	6047.15	11558.95	20605.66
0.3	0.60	368.09	1441.11	3360.83	6539.01	11977.45
0.4	0.10	1186.81	4767.59	11105.95	21136.70	36826.82
0.4	0.15	1064.67	4296.40	10062.06	19276.07	33880.20
0.4	0.30	735.69	3008.63	7164.63	14018.86	25371.21
0.4	0.45	464.84	1923.46	4660.53	9336.08	17492.16
0.4	0.60	254.83	1060.70	2613.34	5371.33	10489.57

Table 7. The present relative efficiency of the proposed estimator $\hat{\pi}_{a1}$ with respect to the direct estimator $\hat{\pi}_{a2}$, i.e. $\text{PRE}(\hat{\pi}_{a1}, \hat{\pi}_{a2})$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.6	0.10	208.54	195.59	185.03	176.09	168.18
0.6	0.15	208.46	195.30	184.41	175.00	166.41
0.6	0.30	208.21	194.39	182.43	171.43	160.34
0.6	0.45	207.95	193.42	180.26	167.31	152.88
0.6	0.60	207.69	192.39	177.84	162.50	143.48
0.7	0.10	459.60	400.58	356.70	322.22	293.55
0.7	0.15	459.21	399.32	354.25	318.18	287.33
0.7	0.30	458.02	395.41	346.46	305.13	266.67
0.7	0.45	456.81	391.28	338.02	290.48	242.39
0.7	0.60	455.56	386.92	328.82	273.91	213.48
0.8	0.10	1168.00	923.94	765.56	652.63	565.52
0.8	0.15	1166.27	919.15	757.19	640.00	547.37
0.8	0.30	1160.98	904.35	731.03	600.00	488.89
0.8	0.45	1155.56	888.89	703.20	556.52	423.53
0.8	0.60	1150.00	872.73	673.53	509.09	350.00

Table 8. The present relative efficiency of the proposed estimator $\hat{\tau}_1$ with respect to the proposed estimator $\hat{\tau}_2$, i.e. $\text{PRE}(\hat{\tau}_1, \hat{\tau}_2)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.6	0.10	155.84	138.02	129.84	125.15	122.10
0.6	0.15	170.09	150.96	141.23	135.34	131.39
0.6	0.30	194.19	177.45	166.92	159.70	154.44
0.6	0.45	206.49	193.96	184.91	178.07	172.72
0.6	0.60	214.04	205.47	198.59	192.95	188.24
0.7	0.10	251.86	197.48	174.56	161.93	153.93
0.7	0.15	300.32	236.32	206.82	189.84	178.81
0.7	0.30	394.56	327.32	289.14	264.52	247.34
0.7	0.45	449.87	393.59	356.26	329.68	309.80
0.7	0.60	486.62	445.05	413.73	389.28	369.66
0.8	0.10	448.61	311.22	257.83	229.44	211.83
0.8	0.15	585.58	407.71	333.71	293.15	267.54
0.8	0.30	901.16	668.74	552.62	482.98	436.57
0.8	0.45	1125.01	897.50	764.03	676.28	614.19
0.8	0.60	1293.88	1104.13	975.00	881.44	810.53

NEW PROCEDURES OF ESTIMATING PROPORTION USING RRT

Table 9. The present relative efficiency of the proposed estimator $\hat{\pi}_{a2}$ with respect to the direct estimator $\hat{\pi}_{a1}$, i.e. $PRE(\hat{\pi}_{a2}, \hat{\pi}_{a1})$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.2	0.10	1168.00	923.94	765.56	652.63	565.52
0.2	0.15	1166.27	919.15	757.19	640.00	547.37
0.2	0.30	1160.98	904.35	731.03	600.00	488.89
0.2	0.45	1155.56	888.89	703.20	556.52	423.53
0.2	0.60	1150.00	872.73	673.53	509.09	350.00
0.3	0.10	459.60	400.58	356.70	322.22	293.55
0.3	0.15	459.21	399.32	354.25	318.18	287.33
0.3	0.30	458.02	395.41	346.46	305.13	266.67
0.3	0.45	456.81	391.28	338.02	290.48	242.39
0.3	0.60	455.56	386.92	328.82	273.91	213.48
0.4	0.10	208.54	195.59	185.03	176.09	168.18
0.4	0.15	208.46	195.30	184.41	175.00	166.41
0.4	0.30	208.21	194.39	182.43	171.43	160.34
0.4	0.45	207.95	193.42	180.26	167.31	152.88
0.4	0.60	207.69	192.39	177.84	162.50	143.48

Table 10. The present relative efficiency of the proposed estimator $\hat{\tau}_2$ with respect to the proposed estimator $\hat{\tau}_1$, i.e. $PRE(\hat{\tau}_2, \hat{\tau}_1)$.

<i>P</i>	<i>T</i>	π				
		0.1	0.2	0.3	0.4	0.5
0.2	0.10	448.61	311.22	257.83	229.44	211.83
0.2	0.15	585.58	407.71	333.71	293.15	267.54
0.2	0.30	901.16	668.74	552.62	482.98	436.57
0.2	0.45	1125.01	897.50	764.03	676.28	614.19
0.2	0.60	1293.88	1104.13	975.00	881.44	810.53
0.3	0.10	251.86	197.48	174.56	161.93	153.93
0.3	0.15	300.32	236.32	206.82	189.84	178.81
0.3	0.30	394.56	327.32	289.14	264.52	247.34
0.3	0.45	449.87	393.59	356.26	329.68	309.80
0.3	0.60	486.62	445.05	413.73	389.28	369.66
0.4	0.10	155.84	138.02	129.84	125.15	122.10
0.4	0.15	170.09	150.96	141.23	135.34	131.39
0.4	0.30	194.19	177.45	166.92	159.70	154.44
0.4	0.45	206.49	193.96	184.91	178.07	172.72
0.4	0.60	214.04	205.47	198.59	192.95	188.24

Variable Selection in Regression using Multilayer Feedforward Network

Tejaswi S. Kamble
Shivaji University
Kolhapur, Maharashtra, India

Dattatraya N. Kashid
Shivaji University
Kolhapur, Maharashtra, India

The selection of relevant variables in the model is one of the important problems in regression analysis. Recently, a few methods were developed based on a model free approach. A multilayer feedforward neural network model was proposed for developing variable selection in regression. A simulation study and real data were used for evaluating the performance of proposed method in the presence of outliers, and multicollinearity.

Keywords: Subset selection, artificial neural network, multilayer feedforward network, full network model and subset network model.

Introduction

The objective of regression analysis is to predict the future value of response variable for the given values of predictor variables. In the regression model, the inclusion of a large number of predictor variables leads to the problems such as i) decrease in prediction accuracy, and ii) increase in cost of the data collection (Miller, 2002). To improve the prediction accuracy of the regression model, one approach is to retain only a subset of relevant predictor variables in the model, and eliminate the irrelevant predictor variables. The problem of choosing an appropriate relevant set from a large number of predictor variables is called subset selection or variable selection in regression.

In traditional regression analysis, the form of the regression model must be first specified, then fitted to the data. However, if a pre-specified form of the model is itself wrong, another model must be used. Searching for a correct model for the given data becomes difficult when complexity is present in the data. A better alternative approach in the above situation would be to estimate a function or model from the data. Such an approach is called Statistical Learning; Artificial

---

Ms. Kamble is a Junior Research Fellow in the Department of Statistics. Email her at tejustat@gmail.com. Dr. Kashid is a Professor in the Department of Statistics. Email him at dnk_stats@unishivaji.ac.in.

Neural Network (ANN) and Support Vector Machine (SVM) are statistical learning techniques.

ANNs have recently received a great deal to attention in many fields of study, such as pattern reorganization, marketing research etc. ANN is important because of its potential use in prediction and classification problems. Usually, ANN is used for prediction when form of the regression model is not specified. In this article, ANN is used for selection of relevant predictor variables in the model.

Mallows's C_p (Mallows, 1973) and S_p statistics (Kashid and Kulkarni, 2002), along with other existing variable selection methods, are suitable under certain assumptions with prior knowledge about the data. When no prior knowledge about the data is available, ANN is an attractive variable selection method (Castellano and Fanelli, 2000), because ANN is a data-based approach. ANN is used in this study for obtaining predicted values of the subset regression model. The criteria C_p and S_p are based on prediction values of subset models. Therefore, we propose modification in C_p and S_p based on predicted values of the ANN model.

Mallows's C_p (Mallows, 1973) is defined by

$$C_p = \frac{RSS_p}{\sigma^2} + (n - 2p) \quad (1)$$

where p is the number of parameters in the subset regression model with $p - 1$ regressors, RSS_p is the residual sum of squares of the subset model, n is the number of data points used for fitting the subset regression model, and σ^2 is replaced by its suitable estimates, usually based on the full model. In this study, the following cases are used.

Case 1

A simulation design proposed by McDonald and Galarneau (1975) is used for introducing multicollinearity in the regressor variables. It is given by

$$X_{ij} = (1 - \rho^2)^{\frac{1}{2}} Z_{ij} + \rho Z_{i(j+1)}, \quad i = 1, 2, \dots, n, j = 1, 2, \dots, J$$

where Z_{ij} are independent standard normal pseudo-random numbers of size n , and ρ^2 is the correlation between any two predictor variables. The response variable Y is generated by using the following regression model with $n = 30$ and $\rho = 0.999$:

VARIABLE SELECTION IN REGRESSION USING MFN

$$Y_i = 1 + 4X_{i1} + 5X_{i2} + 0X_{i3} + \varepsilon_i, i = 1, 2, \dots, 30$$

where $\varepsilon_i \sim N(0,1)$. To identify the degree of multicollinearity, the variance inflation factor (VIF) is used (Montgomery, Peck, and Vining, 2006). For this data, the VIFs for the variables are 339.6, 572.5 and 350.1. These VIFs indicates the presence of severe multicollinearity in the data. We compute the value of the C_p statistic $C_p(M)$ and report the results in Table 1.

Case 2

Data generated in Case 1 is used, and one outlier is introduced by multiplying the actual Y corresponding to the maximum absolute residual by 25. The value of the response variable $Y = 8.2235$ is replaced by $Y = 205.5878$. The value of the C_p statistic $C_p(MO)$ is computed and reported in Table 1.

Case 3

The following nonlinear regression model is generated using the above $X_i, i = 1, 2, 3$ and ε_i which are generated in Case 1. The nonlinear regression model is

$$Y = \exp(1 + 4X_{i1} + 5X_{i2} + 0X_{i3}) + \varepsilon_i, i = 1, 2, \dots, 30$$

The values of the C_p statistic $C_p(NL)$ are computed for the nonlinear regression model and reported in Table 1.

Table 1. Values of $C_p(M)$, $C_p(MO)$, and $C_p(NL)$.

Regressors in subset model	P	$C_p(M)$	$C_p(MO)$	$C_p(NL)$
X_1	2	1.8617	3.0077	2.0726
X_2	2	2.2565	2.2510	1.0605
X_3	2	3.2585	1.9152	2.3498
X_1X_2	3	2.2237	2.8740	2.0059
X_1X_3	3	3.8518	3.2340	3.8492
X_2X_3	3	4.1730	3.4448	3.0179
$X_1X_2X_3$	4	4.0000	4.0000	4.0000

As seen in Table 1, the criterion C_p selects the wrong subset models for all the above-cited cases. The statistic fails to select the correct model in the presence of a) multicollinearity alone, b) both multicollinearity and outlier, and c)

nonlinear regression, because OLS estimation does not perform well in each case. Consequently, variable selection methods based on OLS estimator fail to select the correct model.

Regression Model and Neural Network Model

In general, the regression model is defined as

$$\mathbf{Y} = f(X, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \quad (2)$$

where f is any function of predictor variables $X_1, X_2, \dots, X_{k-1}$ and unknown regression coefficients $\boldsymbol{\beta}$. If f is a non-linear function, then regression parameters are estimated by using nonlinear least squares method (or some other method). If f is linear, the regression model can be expressed as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (3)$$

where $\mathbf{Y}$ is an $n \times 1$ vector of response variables, $\mathbf{X}$ is a matrix of order $n \times k$ with 1's in the first column, $\boldsymbol{\beta}$ is a $k \times 1$ vector of regression coefficients and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of random errors which are independent and identically distributed $N(0, \sigma^2 \mathbf{I})$. The least squares estimator of $\boldsymbol{\beta}$ is given by (Montgomery et al., 2006)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}$$

The predicted value of the regression model is obtained by the fitted equation

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The prediction accuracy of the regression model depends on the selection of an appropriate model, which means the form of the function (f) must be specified before the regression analysis. If form of the model is not known, then one of the most appropriate alternative methods to handle this situation is artificial neural network.

Multilayer Feedforward Network (MFN)

The MFN can approximate any measurable function to any desired degree of accuracy (Hornik, Stinchcombe, and White, 1989). This MFN model consists of an input layer, an output layer, and one or more hidden layer(s). We represent the architecture of MFN with one hidden layer consisting of J hidden nodes, and a single node in an output layer, as shown in Figure 1. A vector $\mathbf{X} = [X_0, X_1, \dots, X_{k-1}]'$ is the vector of k units in the input layer and $\mathbf{Y}$ is the output of the network.

Figure 1. Multilayer feedforward network

From Figure 1, each input signal is connected to each node in the hidden layer with weight w_{jm} , $m = 0, 1, 2, 3, \dots, k-1$, $j = 1, 2, \dots, J$, and hidden nodes are connected to a node in the output layer with weight v_j , $j = 1, 2, \dots, J$. The final output Y_i for the i^{th} data point is given by

$$Y_i = g_2 \left(\sum_{j=1}^J v_j g_1 \left(\sum_{m=0}^{k-1} w_{jm} X_{im} \right) \right) \quad i = 1, 2, \dots, n$$

where g_1 and g_2 denote activation functions used in the hidden layer and output layer respectively; it is not necessary that g_1 and g_2 are the same activation functions. The above network model can be written as

$$\mathbf{Y} = f(X, \boldsymbol{\beta}) \quad (4)$$

where $\boldsymbol{\beta} = (v_1, \dots, v_J, \mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_{k-1})$, $\mathbf{w}_m = (w_{1m}, w_{2m}, \dots, w_{Jm})$, $m = 0, 1, 2, \dots, k-1$ and $f(X, \boldsymbol{\beta})$ is a nonlinear function of the inputs $X_0, X_1, X_2, \dots, X_{k-1}$ and the weight vector $\boldsymbol{\beta}$. If we add an error term in the above model (4), then it becomes a regression model as in Equation 2, where $\boldsymbol{\varepsilon}$ is the random error.

The next step in ANN modeling is training the network. The purpose of training the network is to obtain weights in a neural network model using the training data. Various training methods or algorithms are available in the literature. The robust back-propagation method (see Kasko, 1992) is one such. First, two types of MFN models must be defined, namely the full MFN model and the subset MFN model, for proposing modification in C_p and S_p statistics.

Full MFN and subset MFN model

A full MFN model is constructed with input units $X_1, X_2, \dots, X_{k-1}$ and bias node $X_0 = -1$. The MFN model in Equation 4 is a full MFN model. The network weights are obtained by training the network and the network output vector based on a full MFN model, as

$$\hat{\mathbf{Y}} = f(X, \hat{\boldsymbol{\beta}}) \quad (5)$$

where $\hat{\boldsymbol{\beta}}$ is the estimated weight vector.

A subset MFN model is constructed with a subset of input units $X_A = (X_0, X_1, X_2, \dots, X_{p-1})'$ of size $p (p \leq k)$ in the input layer. The subset network model is given by

$$\mathbf{Y} = f(X_A, \boldsymbol{\beta}_A) \quad (6)$$

where X and $\boldsymbol{\beta}$ are partitioned as $X = [X_A : X_B]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_A : \boldsymbol{\beta}_B]$. Similarly, the network output vector based on subset MFN model is

$$\hat{\mathbf{Y}} = f(X_A, \hat{\boldsymbol{\beta}}_A) \quad (7)$$

where $\hat{\boldsymbol{\beta}}_A$ is the estimated weight vector.

To implement the training procedure using network training algorithm, we need to select the number of hidden layers in the MFN and the number of hidden nodes in that hidden layer. This is discussed in the next section.

Selection of Hidden Layer and Hidden Nodes

The selection of learning rate parameter, initial weights and number of hidden layers in the MFN model and the number of hidden nodes in each hidden layer is an important task. The number of hidden layers is determined first. The network begins as a one-hidden-layer network (Lawrence, 1994). If the one-hidden-layer MFN network does not sufficient for training the network, then more hidden layers are added. In the MFN model, theoretically a single hidden layer is sufficient, because any continuous function defined on a compact set in R^n can be approximated by a multilayer ANN with one hidden layer with sigmoid activation function (Cybenko, 1989). Based on this result, we consider the single hidden layer MFN model with sigmoid activation function.

The choice of number of hidden neurons in the hidden layer is also a considerable problem, and it depends on the data. Research has proposed various methods for selection of hidden nodes in the hidden layer (see Chang-Xue, Zhi-Guang and Kusiak, 2005), as follows:

- $H_1 = 2I + 1$ (Hecht-Nelson, 1987)
- $H_2 = (I + O)/2$ (Lawrence and Fredrickson, 1998)
- $n/10 - I - O \leq H_3 \leq n/2 - I - O$ (Lawrence and Fredrickson, 1998)
- $H_4 = I \log_2 n$ (Marchandani and Cao, 1989)
- $H_5 = O(I + 1)$ (Lipmann, 1987)

Here, I is the number of inputs, O is the number of output neurons, and n is the number of training data points.

Variable Selection Methods and Proposed Methods

In the classical linear regression, several variable selection procedures have been suggested by the researchers. Most methods are based on least squares (LS) parameter estimation procedure. The variable selection methods based on LS estimates of β fail to select the correct subset model in the presence of outlier, multicollinearity, or nonlinear relationship between Y and X . Here, we modified existing subset selection methods using MFN model for prediction.

It is demonstrated that the Mallows's C_p statistic does not work well when assumptions are violated. Researchers have suggested some other methods for variable selection (see Ronchetti and Staudte, 1994; Sommer and Huggins, 1996). Also Kashid and Kulkarni (2002) have suggested a more general criterion, the S_p statistic for variable selection in cases of clean and outlier data. It can be defined as

$$S_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} + (k - 2p) \quad (8)$$

where $\hat{Y}_{ik}$ is the predicted value of the full model, $\hat{Y}_{ip}$ is the predicted value of the subset model based on M-estimator of the regression parameters, and k and p are the number of parameters in the full and subset model respectively. The σ^2 is replaced by its suitable estimates, which usually consists of the full model.

The subset selection procedure is same for both the methods. The S_p statistic is equivalent to the C_p statistic when LS method is used for estimating regression coefficients. The following suggests modification in both criteria using the complicity measure.

MC_p and MS_p Criteria

In a modified version of the C_p and S_p statistics, the network output (estimated values of response $\mathbf{Y}$) is obtained by using the single hidden layer with a single output MFN model.

The network outputs $\hat{Y}_{ik} = f(\mathbf{X}_i, \hat{\boldsymbol{\beta}})$ and $\hat{Y}_{ip} = f(\mathbf{X}_{iA}, \hat{\boldsymbol{\beta}}_A)$ denote outputs based on full MFN and subset MFN model, respectively. The residual sum of squares for the full and subset network models are defined as

$$RSS_k = \sum_{i=1}^n (Y_i - \hat{Y}_{ik})^2, \text{ and}$$

$$RSS_p = \sum_{i=1}^n (Y_i - \hat{Y}_{ip})^2$$

The modified version of C_p and S_p are denoted as MC_p and MS_p . They are defined by

$$MC_p = \frac{RSS_p}{\sigma^2} + C(n, p), \text{ and} \quad (9)$$

$$MS_p = \frac{\sum_{i=1}^n (\hat{Y}_{ik} - \hat{Y}_{ip})^2}{\sigma^2} + C(n, p) \quad (10)$$

where n is the number of data points and p is the number of inputs including bias node (X_0). $\hat{Y}_{ik}$ and $\hat{Y}_{ip}$ are the predicted values of Y based on the full and subset MFN models, respectively, $C(n, p)$ is the penalty term, and σ^2 is replaced by its suitable estimate if it is unknown. The motivation for proposing modified versions of C_p and S_p are as follows.

In criterion MC_p , we use two types of measures. The first term measures the discrepancy between the desired output and network output based on the subset MFN model. The smaller this value is, the closer to the desired output it is; the smallest value of this measure is smallest for the full model. Therefore, it is difficult to select the correct model by minimizing criterion. So, we add a complicity measure called the penalty function, comprised of only p , only n , or both n and p .

In the second criterion MS_p , we use sum of squared difference between network output of the full and subset MFN models. The smallest value indicates that a prediction based on the subset MFN model is as accurate as the full MFN model. When full MFN model is itself the correct model, this value is zero. It is difficult to select the correct model using the minimizing criterion. Therefore we added the penalty function similar to criterion defined in (9) and used the same logic for the selection of subset. The selection procedure for both methods is as follows.

- Step I: Compute the MC_p for all possible subsets.
- Step II: Select the subset corresponding to the minimum value of MC_p .
Use the same procedure for MS_p .

Choice of Estimator of σ^2

An estimator of σ^2 is required to implement the MC_p and MS_p criteria. In the literature of regression, various estimators of σ^2 are available. What follows are estimators of σ^2 used in MC_p and MS_p based on full network output, and a study of the effect of these estimators on the value of MC_p and MS_p .

1.
$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_{ik})^2}{n - k}$$
2.
$$\hat{\sigma}_2^2 = \left(1.4826 \text{median} |r_i - \text{median}(r_i)|\right)^2$$
3.
$$\hat{\sigma}_3^2 = \left(1.4826 \text{median} |r_i|\right)^2$$

where n is the number of data points, k is the number of inputs in the full MFN model including bias node $r_i = Y_i - \hat{Y}_{ik}$, and $\hat{Y}_{ik}$ is the network output for the i^{th} data point based on the full MFN model.

Performances of MC_p and MS_p

To evaluate the performance of MC_p and MS_p , we have used single hidden layer MFN model and robust back-propagation training method with sigmoid activation function in the hidden layer and output layer. In robust back-propagation, we use an error suppressor function $s(e)$ by replacing the scalar squared error e (Kasko, 1992), because $s(e) = e^2$ is not robust. The following error suppressor functions are used in this study.

1. $E_1 = s(e) = \max(-c, \min(c, e))$ (Huber function)
(where $c = 1.345$ is bending constant)
2. $E_2 = s(e) = 2e/(1+e^2)$ (Cauchy function)
3. $E_2 = s(e) = \tanh(e/2)$ (Hyperbolic tangent function)

The learning rate parameter (η) is selected by trial and error, and the number of hidden nodes in hidden layer is selected using the selection methods given earlier. The following seven penalty functions are used for computing MS_p and MC_p ; some are available in the literature (Sakate and Kashid, 2014).

1. $P_1 = 2p$

2. $P_2 = p \log(n+2)$
3. $P_3 = 2p + \frac{2(p+1)(p+2)}{n-p-2}$
4. $P_4 = p(\log n + 1)$
5. $P_5 = \frac{2pn}{n-p-1}$
6. $P_6 = 2p + \frac{2p(p+1)}{n-p-1}$
7. $P_7 = p \log n$

The performance of the proposed methods is measured for different combinations of penalty functions $(P_l) l = 1, 2, \dots, 7$, selection methods of hidden nodes in the hidden layer $(H_m) m = 1, 2, \dots, 5$, and error suppressor functions $(E_o) o = 1, 2, 3$; these are denoted by (P_l, H_m, E_o) . Three simulation designs are used for the evaluation of the performance of MS_p and MC_p .

Simulation Design A

The performance of proposed modified versions of $S_p(MS_p)$ and $C_p(MC_p)$ are evaluated using the following models with two error distributions.

Model I: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$, where $\beta = (1, 5, 10, 0)$,

Model II: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$, where $\beta = (1, 5, 10, 0, 0)$

The regressor variables were generated from $U(0,1)$ and the error term was generated from $N(0,1)$ and Laplace $(0,1)$. The response variable Y was generated using Models I and II for sample sizes 20 and 30, respectively. This experiment is repeated 100 times and ability of these methods to select the correct model is measured using learning parameter $(\eta) = 0.1$ and $\hat{\sigma}_1^2$. The results are reported in Tables 2 through 5.

Table 2. Model selection ability of MS_p and MC_p in 100 replications for Model I of size 20

Error distribution	Error suppressor function	P_n	H_1		H_2		H_3		H_4		H_5	
			MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	79	66	84	77	72	75	73	64	77	71
		P_2	86	81	92	82	81	87	84	77	87	84
		P_3	88	86	94	90	90	92	89	86	93	89
		P_4	88	85	94	88	88	90	87	81	90	87
		P_5	86	81	92	85	82	87	85	79	88	85
		P_6	86	81	92	85	82	87	85	79	88	85
		P_7	85	79	92	82	79	87	82	77	87	84
	Cauchy	P_1	78	58	77	32	76	52	67	57	63	69
		P_2	91	71	85	35	83	72	79	68	80	76
		P_3	93	79	85	34	86	77	87	80	84	83
		P_4	92	74	85	36	84	77	84	74	83	81
		P_5	91	71	85	36	83	72	79	69	82	76
		P_6	91	71	85	36	83	72	79	69	82	76
		P_7	91	70	85	35	82	72	79	66	79	75
	Hyperbolic Tangent	P_1	79	66	74	77	75	79	75	79	77	83
		P_2	86	81	86	84	85	87	85	87	86	91
		P_3	88	86	91	89	87	90	87	90	92	91
		P_4	88	85	88	86	86	89	86	89	89	91
		P_5	86	81	86	84	85	88	85	88	87	91
		P_6	86	81	86	84	85	88	85	88	87	91
		P_7	85	79	85	84	85	87	85	87	85	91
Laplace	Huber	P_1	69	67	75	66	75	69	77	34	78	66
		P_2	83	81	86	80	87	73	89	36	79	79
		P_3	86	86	91	84	89	80	94	35	80	81
		P_4	87	83	88	82	89	76	93	36	81	81
		P_5	84	81	86	80	87	73	91	36	80	79
		P_6	84	81	86	80	87	73	91	36	80	79
		P_7	81	81	86	77	85	73	88	35	79	79
	Cauchy	P_1	74	54	77	52	68	67	70	51	71	62
		P_2	83	75	81	60	80	77	80	66	78	74
		P_3	86	85	86	67	84	80	85	76	80	81
		P_4	86	84	84	65	82	79	84	72	79	78
		P_5	84	77	82	60	80	77	82	67	78	74
		P_6	84	77	82	60	80	77	82	67	78	74
		P_7	83	74	80	60	79	77	79	65	75	73
	Hyperbolic Tangent	P_1	70	67	76	69	85	76	85	76	82	63
		P_2	83	81	82	82	90	85	90	85	88	75
		P_3	86	86	87	88	92	89	92	89	93	75
		P_4	87	84	86	87	92	88	92	88	93	78
		P_5	84	81	83	83	90	85	90	85	88	76
		P_6	84	81	83	83	90	85	90	85	88	76
		P_7	82	81	82	82	90	84	90	84	87	74

VARIABLE SELECTION IN REGRESSION USING MFN

Table 3. Model selection ability of MS_p and MC_p in 100 replications for Model I of size 30

Error distribution	Error suppressor function	P_n	H_1		H_2		H_3		H_4		H_5	
			MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	78	72	78	74	71	69	76	62	74	72
		P_2	89	81	89	88	83	85	90	74	90	92
		P_3	93	87	92	92	92	87	94	96	92	94
		P_4	88	77	84	84	78	82	92	72	85	80
		P_5	87	77	82	82	77	79	92	66	80	79
		P_6	87	77	82	82	77	79	92	66	80	78
		P_7	89	81	88	88	83	85	90	74	88	92
	Cauchy	P_1	72	59	74	71	77	59	76	52	70	50
		P_2	85	73	81	88	84	74	86	68	86	76
		P_3	94	82	87	93	88	81	94	80	94	80
		P_4	80	66	83	83	83	69	84	62	80	68
		P_5	79	65	82	79	81	68	84	60	80	66
		P_6	79	65	82	79	81	68	84	61	80	66
		P_7	84	73	81	88	84	74	86	68	86	68
	Hyperbolic Tangent	P_1	83	74	82	71	78	74	74	62	78	76
		P_2	89	82	93	88	92	87	82	72	90	88
		P_3	94	87	96	92	94	91	86	68	96	92
		P_4	85	81	91	81	88	83	86	72	84	83
		P_5	85	81	88	79	86	82	82	70	85	82
		P_6	85	81	88	79	86	82	82	71	84	82
		P_7	88	92	93	88	91	86	82	74	90	86
Laplace	Huber	P_1	73	56	77	70	72	54	80	58	78	62
		P_2	82	75	91	85	91	80	80	78	88	80
		P_3	89	81	92	87	90	84	86	86	90	86
		P_4	82	70	85	81	82	75	81	70	90	76
		P_5	81	66	84	77	82	72	81	64	91	72
		P_6	81	66	84	77	82	73	81	65	84	72
		P_7	82	74	91	85	88	80	80	72	88	80
	Cauchy	P_1	62	33	74	47	77	66	76	56	77	60
		P_2	78	43	83	66	86	78	86	66	85	76
		P_3	87	58	87	73	90	80	92	80	87	84
		P_4	75	40	81	58	84	77	80	62	84	70
		P_5	73	38	80	56	82	75	78	62	84	66
		P_6	73	38	80	56	82	75	78	62	84	66
		P_7	77	43	83	64	86	78	86	66	84	74
	Hyperbolic Tangent	P_1	72	77	72	71	78	68	78	60	82	50
		P_2	85	90	89	84	85	86	82	78	96	76
		P_3	88	93	91	89	90	88	86	86	97	84
		P_4	82	87	84	83	84	83	78	78	94	70
		P_5	82	86	83	80	82	80	78	78	94	62
		P_6	82	86	83	80	82	80	78	78	94	62
		P_7	84	90	89	84	85	87	80	80	98	76

Table 4. Model selection ability of MS_p and MC_p in 100 replications for Model II of size 20

Error distribution	Error suppressor function	P_n	H_1		H_2		H_3		H_4		H_5	
			MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	60	33	60	43	62	50	62	38	68	60
		P_2	79	53	77	59	72	72	76	60	74	72
		P_3	85	68	83	78	82	82	85	72	78	85
		P_4	82	64	83	65	83	78	80	78	76	80
		P_5	80	57	79	60	72	74	76	64	74	76
		P_6	80	57	79	60	72	74	76	64	74	76
		P_7	77	53	76	59	72	70	76	58	74	72
	Cauchy	P_1	54	40	51	24	60	22	48	32	60	43
		P_2	68	40	72	46	70	38	76	49	70	56
		P_3	72	43	80	68	82	50	80	56	76	65
		P_4	71	45	75	64	80	46	80	52	76	63
		P_5	69	51	73	46	70	38	78	49	78	58
		P_6	69	63	73	46	70	38	78	49	78	58
		P_7	66	50	71	42	68	38	74	49	70	56
	Hyperbolic Tangent	P_1	63	42	69	60	50	50	61	44	68	70
		P_2	74	72	78	72	68	74	88	65	84	84
		P_3	82	85	82	78	74	82	88	78	94	86
		P_4	79	83	82	74	74	78	88	78	90	86
		P_5	75	76	78	74	70	78	88	78	89	85
		P_6	75	76	79	74	70	76	88	68	88	84
		P_7	72	70	79	74	66	70	89	68	80	84
Laplace	Huber	P_1	40	44	54	32	56	35	68	48	41	40
		P_2	62	58	68	52	67	56	76	72	62	60
		P_3	76	66	88	78	74	75	74	65	70	74
		P_4	70	65	72	63	76	73	82	76	64	70
		P_5	65	59	68	52	66	60	76	72	60	60
		P_6	65	59	68	52	66	60	76	72	61	60
		P_7	58	58	67	50	66	54	76	70	60	56
	Cauchy	P_1	59	29	50	32	52	32	44	22	44	49
		P_2	61	40	64	48	74	50	56	45	64	62
		P_3	64	53	65	56	78	60	58	53	73	72
		P_4	65	50	64	52	76	58	56	52	67	68
		P_5	64	43	65	48	74	50	56	48	64	64
		P_6	64	43	65	48	75	50	56	48	64	64
		P_7	61	40	62	44	75	46	54	43	62	58
	Hyperbolic Tangent	P_1	54	44	58	44	56	35	52	38	60	60
		P_2	78	60	78	70	67	57	60	53	74	72
		P_3	74	66	84	76	74	74	61	56	87	81
		P_4	74	66	83	76	78	76	62	54	83	80
		P_5	72	60	78	70	66	60	61	52	74	74
		P_6	72	60	78	70	66	60	61	52	74	74
		P_7	70	60	78	78	66	54	61	50	72	76

VARIABLE SELECTION IN REGRESSION USING MFN

Table 5. Model selection ability of MS_p and MC_p in 100 replications for Model II of size 30

Error distribution	Error suppressor function	P_n	H_1		H_2		H_3		H_4		H_5	
			MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	Huber	P_1	69	36	64	55	64	30	72	46	66	46
		P_2	82	77	83	64	76	60	84	70	84	66
		P_3	83	87	86	73	78	80	86	76	84	88
		P_4	80	66	80	63	76	43	82	64	80	64
		P_5	78	85	72	60	74	40	78	60	78	62
		P_6	78	58	72	61	74	39	78	60	77	62
		P_7	83	77	82	64	75	60	84	70	80	66
	Cauchy	P_1	45	25	51	44	52	30	52	23	44	34
		P_2	68	58	65	68	71	60	72	40	62	52
		P_3	79	68	74	74	78	66	79	58	78	62
		P_4	56	51	64	64	68	44	66	32	54	42
		P_5	57	38	64	64	66	45	65	30	46	42
		P_6	57	38	64	64	66	44	64	30	46	42
		P_7	66	54	64	68	70	58	65	40	62	52
	Hyperbolic Tangent	P_1	68	36	70	57	52	53	72	44	56	35
		P_2	82	76	80	78	70	69	84	72	76	62
		P_3	82	86	80	86	80	82	86	76	86	80
		P_4	80	66	78	72	70	74	81	64	68	52
		P_5	76	60	76	68	66	69	80	62	68	48
		P_6	76	60	76	69	66	69	79	62	68	48
		P_7	82	76	81	76	70	69	84	70	32	63
Laplace	Huber	P_1	56	36	54	48	52	56	48	52	52	36
		P_2	86	50	72	70	74	84	70	74	76	70
		P_3	92	54	78	74	84	92	74	80	84	70
		P_4	74	46	66	64	69	80	66	72	70	50
		P_5	74	46	64	64	62	70	64	72	66	46
		P_6	74	46	63	64	62	70	64	72	66	46
		P_7	86	50	72	68	74	84	68	74	76	70
	Cauchy	P_1	32	36	60	24	50	34	40	21	36	21
		P_2	52	60	80	42	60	62	74	45	56	48
		P_3	64	74	86	48	74	70	84	56	64	60
		P_4	40	54	68	32	52	54	62	32	45	36
		P_5	40	52	66	30	50	48	56	28	42	32
		P_6	40	52	66	31	50	48	56	28	42	33
		P_7	48	60	80	40	61	62	72	42	42	42
	Hyperbolic Tangent	P_1	66	44	52	46	50	81	60	46	52	36
		P_2	80	72	80	66	72	68	81	70	79	64
		P_3	84	80	84	79	76	80	86	79	86	82
		P_4	74	66	71	62	74	68	81	66	60	56
		P_5	72	30	64	56	72	68	75	62	60	48
		P_6	72	61	64	56	72	68	76	62	60	48
		P_7	80	70	76	66	72	68	83	70	74	74

From Tables 2 through 5, it can be observed that the overall performance of the MS_p statistic is better than the MC_p statistic. The performance of penalties P_2 through P_7 is better than penalty P_1 , with H_1 through H_5 , for Models I and II. Based on these simulations, it is recommended that any hidden node selection method be used with penalty P_2 through P_7 and Huber or Hyperbolic Tangent error suppressor function.

Simulation Design B

The experiment was repeated 100 times using the simulation design A. The performance of MS_p and MC_p were compared with Mallows's C_p for Models I and II with sample sizes of 20 and 30. MS_p and MC_p were computed using (P_3, H_1, E_1) , and learning parameters $(\eta) = 0.1$ and $\hat{\sigma}_1^2$. The results are reported in Table 6.

Table 6. Model selection ability of correct model for 100 repetitions

Error Distribution	Sample sizes	Model I			Model II		
		MS_p	MC_p	C_p	MS_p	MC_p	C_p
Normal	20	94	90	82	83	78	76
	30	92	92	79	86	73	70
Laplace	20	91	84	81	88	78	77
	30	92	87	84	78	74	75

From Table 6, it is clear that the model selection ability of MS_p and MC_p is better than C_p (based on LS estimates) for sample sizes 20 and 30 for both error distributions. The model selection ability of MS_p is uniformly larger than that of MC_p or C_p .

Simulation Design C

Three further models based on MFN are used to evaluate the performance of MS_p and MC_p :

$$\text{Model III: } Y = \sqrt{\beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2} + \varepsilon,$$

$$\text{Model IV: } Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2 + \varepsilon,$$

$$\text{Model V: } Y = e^{\beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_3 X_3^2 + \beta_4 X_4^2} + \varepsilon,$$

where $\beta = (1, 5, 10, 0, 0)$.

In this simulation, $X_i = (i = 1, 2, 3, 4)$ were generated from $U(0, 1)$ and error was generated from $N(0, 1)$ and $\text{Laplace}(0, 1)$. The response variable Y was generated using Models III, IV and V. MS_p and MC_p were computed using $(P_1 -$

VARIABLE SELECTION IN REGRESSION USING MFN

P_7, H_1, E_1), learning parameters $(\eta) = 0.1$ and $\hat{\sigma}_1^2$. The ability of these methods to select the correct model over 100 replications is reported in Table 7.

Table 7. Correct model selection ability over 100 replications

Error distribution	P_n	Model III				Model IV				Model V			
		$n = 20$		$n = 30$		$n = 20$		$n = 30$		$n = 20$		$n = 30$	
		MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p	MS_p	MC_p
Normal	P_1	50	40	78	25	71	57	89	65	04	07	72	76
	P_2	55	35	89	48	78	70	91	73	05	06	90	91
	P_3	55	24	93	58	83	78	88	60	04	07	90	95
	P_4	60	38	80	34	80	76	82	56	05	07	91	85
	P_5	54	37	77	32	79	72	83	56	05	07	83	82
	P_6	55	40	77	35	79	72	85	65	05	06	89	82
	P_7	54	34	90	42	76	69	90	70	05	06	75	90
Laplace	P_1	20	16	60	40	15	16	89	70	07	05	89	19
	P_2	21	14	80	66	12	14	93	80	07	04	99	18
	P_3	25	15	86	80	7	11	82	65	06	04	100	13
	P_4	22	14	75	56	12	15	80	52	05	03	96	10
	P_5	20	14	75	50	13	16	80	52	05	04	90	16
	P_6	20	15	75	50	13	16	90	70	08	05	90	16
	P_7	18	14	80	64	13	14	91	72	04	06	99	14

From Table 7, it is clear that performance of MS_p is better than MC_p for all models and sample size 30. The performance of both criteria MS_p and MC_p is very poor for all models when error distribution is Laplace for small samples: the sample size must be moderate to large for selection of relevant variables when regression model is nonlinear.

Performance of MC_p and MS_p in the presence of multicollinearity and outlier

The performance of MS_p and MC_p is studied using the Hald data (Montgomery et al, 2006). The variance inflation factors (VIF) corresponding to each term are 38.5, 254.4, 46.9, and 282.5. The VIF values indicate that multicollinearity exists in the data. Consider the following cases:

- Case I: Data with multicollinearity (original data)
- Case II: Data with multicollinearity and single outlier ($Y_6 = 109.2$ is replaced by 150)
- Case III: Data with multicollinearity and two outliers ($Y_2 = 73.4$ and $Y_6 = 109.2$ are replaced by 150 and 200 respectively)

MS_p and MC_p was computed for all possible subset models with different penalty functions and estimators of σ^2 . The selected subset model, by various combinations of $(P_l, \hat{\sigma}_s^2)$, $l = 1, 2, \dots, 7$, $s = 1, 2, 3$ is reported in Table 8. For training the network, the simulation employs the Huber error suppressor function, number of hidden neurons H_1 , and learning parameter $(\eta) = 0.1$. The results are reported in Table 8.

Table 8. Selected subset by MS_p and MC_p for Cases I – III

Statistic	P_n	Case I			Case II			Case III		
		σ_1^2	σ_2^2	σ_3^2	σ_1^2	σ_2^2	σ_3^2	σ_1^2	σ_2^2	σ_3^2
MS_p	P_1	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2
	P_2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2
	P_3	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2
	P_4	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X2	X1X2	X1X2
	P_5	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X2	X1X2	X1X2
	P_6	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X2	X1X2	X1X2
	P_7	X1X2	X1X2	X1X2	X1X2	X1X2	X1X2	X2	X1X2	X1X2
MC_p	P_1	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X1X2	X1X4	X1X4
	P_2	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X1X2	X1X4	X1X4
	P_3	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X1X2	X1X4	X1X4
	P_4	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X2	X1X4	X1X4
	P_5	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X2	X1X4	X1X4
	P_6	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X2	X1X4	X1X4
	P_7	X1X4	X1X4	X1X4	X1X4	X1X4	X1X4	X2	X1X4	X1X4

This data is analyzed in the connection of multicollinearity and outlier (see Ronchetti and Staudte, 1994; Sommer and Huggins, 1996; and Kashid and Kulkarni, 2002). They have suggested $\{X_1, X_2\}$ is the best subset model for clean data and outlier data. The MS_p statistic selects the same subset model for all combinations of $(P_l, \hat{\sigma}_s^2)$, $l = 1, 2, \dots, 7$, $s = 1, 2, 3$, for Case I and II. In Case III, MS_p fails to select correct model for penalty $P_4 - P_7$ with $\hat{\sigma}_1^2$. Conclusion: the MS_p statistic performs better than MC_p for all cases with all penalty functions and estimators of σ^2 , excluding few cases.

Conclusion

The proposed modified methods are model-free. It is clear that the performance of proposed MS_p statistic is better than classical regression methods in the presence of multicollinearity, outlier, or both simultaneously. The MS_p statistic selects the correct model in cases of nonlinear model for moderate to large sample sizes. From the simulation study, it can be observed that MFN is useful when there is no idea about the functional relationship between response and predictor variables. The MS_p statistic is also useful for selection of inputs from a large set of inputs in a network model, in order to find which network output is closest to the desired output.

Acknowledgements

This research was partially funded by the University Grant Commission, New Delhi, India, under the Rajiv Gandhi National Fellowship scheme vide letter number F.14-2(SC)/2010(SA-III).

References

- Chang-Xue, J. F., Zhi-Guang, Yu. and Kusiak, A. (2006) Selection and validation of predictive regression and neural network models based on designed experiment. *IIE Transactions*, 38(1), 13-23. doi: [10.1080/07408170500346378](https://doi.org/10.1080/07408170500346378)
- Cybenko, G. (1989) Approximation by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4), 303-314. doi: [10.1007/BF02551274](https://doi.org/10.1007/BF02551274)
- Castellano G. and Fanelli A. M. (2000) Variable selection using neural network models. *Neurocomputing*, 31(1-4), 1-13. doi: [10.1016/S0925-2312\(99\)00146-0](https://doi.org/10.1016/S0925-2312(99)00146-0)
- Hecht-Nelson, R. (1987) Kolmogorov's mapping neural network existence theorem. In *Proceedings of the IEEE International Conference on Neural Networks 111*. New York: IEEE Press, pp. 11-14.
- Hornik, K., Stinchcombe, M. and White, H. (1989) Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 359-366.
- Kashid, D. N. and Kulkarni, S. R. (2002) A more general criterion for subset selection in multiple linear regressions. *Communication in Statistics-Theory & Method*, 31(5), 795-811. doi: [10.1081/STA-120003653](https://doi.org/10.1081/STA-120003653)

- Kasko, B. (1992) *Neural networks and fuzzy systems: a dynamic systems approach to machine intelligence*. Englewood Cliffs, N.J.: Prentice-Hall, Inc.
- Lawrence, J. (1994) *Introduction to neural networks: design theory and applications*, 6th Ed. Nevada City, CA: California Scientific Software.
- Lawrence, J. and Fredrickson, J. (1998) *Brain Maker user's guide and reference manual*. Nevada City, CA: California Scientific Software.
- Lippmann, R. P. (1987) An introduction to computing with neural nets. *IEEE Acoustics, Speech and Signal Processing Magazine*, 4(2), 4–22. doi: [10.1109/MASSP.1987.1165576](https://doi.org/10.1109/MASSP.1987.1165576)
- Mallows, C. L. (1973) Some comments on C_p . *Technometrics*, 15(4), 661–675. doi: [10.1080/00401706.1973.10489103](https://doi.org/10.1080/00401706.1973.10489103)
- Marchandani, G. and Cao, W. (1989) On hidden nodes for neural nets. *IEEE Transactions on Circuits and Systems*, 36(5), 661–664. doi: [10.1109/31.31313](https://doi.org/10.1109/31.31313)
- McDonald, G. C. and Galarneau, D. I. (1975) A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350), 407–412. doi: [10.1080/01621459.1975.10479882](https://doi.org/10.1080/01621459.1975.10479882)
- Miller, A. J. (2002) *Subset selection in regression*. London: Chapman and Hall.
- Montgomery, D., Peck, E. and Vining, G. (2006) *Introduction to linear regression analysis*. New York: John Wiley and Sons Inc.
- Ronchetti, E. M. and Staudte, R. G. (1994) A robust version of Mallows's C_p . *Journal of the American Statistical Association*, 89(426), 550–559. doi: [10.1080/01621459.1994.10476780](https://doi.org/10.1080/01621459.1994.10476780)
- Sakate D. M. and Kashid D. N. (2014). A deviance-based criterion for model selection in GLM. *Statistics: A Journal of Theoretical and Applied Statistics*, 48(1), 34–48. doi: [10.1080/02331888.2012.708035](https://doi.org/10.1080/02331888.2012.708035)
- Sommer S. and Huggins R. M. (1996). Variable selection using the Wald test and a robust C_p . *Journal of the Royal Statistical Society C: Applied Statistics*, 45(1), 15–29. doi: [10.2307/2986219](https://doi.org/10.2307/2986219)

Principal Component Preliminary Test Estimator in the Linear Regression Model

Sivarajah Arumairajan
University of Jaffna
Sri Lanka

Pushpakanthie Wijekoon
University of Peradeniya
Sri Lanka

A Preliminary Test Estimator is introduced based on Principal Component Regression Estimator defined in the linear regression model when the stochastic restrictions are available in addition to the sample information, and when the explanatory variables are multicollinear. It is further developed as a large sample preliminary test estimator by using Wald (WA), Likelihood Ratio (LR), and Lagrangian Multiplier (LM) tests. Stochastic properties of this estimator based on F test as well as WA, LR, and LM tests are derived, and the performance of the estimator is compared using WA, LR, and LM tests with respect to Mean Square Error Matrix (MSEM). A Monte Carlo simulation is carried out to illustrate the theoretical findings.

Keywords: Principal Component Regression, Preliminary Test Estimator, Wald Test, Likelihood Ratio Test, Lagrangian Multiplier Test, Mean Square Error Matrix

---

Introduction

Instead of using the Ordinary Least Square Estimator (OLSE), some biased estimation procedures were developed in the literature to combat the multicollinearity problem in the linear regression model. Some of these are namely the Principal Component Regression Estimator (PCRE) (Massy, 1965), Ridge Estimator (RE) (Hoerl & Kennard, 1970) and Liu Estimator (LE) (Liu, 1993). Another way of solving the multicollinearity problem is to consider parameter estimation with some additional information on the unknown parameters such as the exact or stochastic restrictions. By adding exact restrictions to a sample model, the resulting Restricted Least Squares Estimator (RLSE) might again be better in the mean square error sense than the OLSE. By grafting the ridge regression philosophy into the RLSE, the Restricted Ridge

---

Sivarajah Arumairajan is a lecturer in the Department of Mathematics and Statistics. Email at: arumais@gmail.com. Pushpakanthie Wijekoon is a Professor in the Department of Statistics & Computer Science.

Estimator (RRE) has been proposed by Sarkar (1992). As done by Liu (1993), Kaçiranlar, Sakallioğlu, Akdeniz, Styan, & Werner (1999) proposed a biased estimator called Restricted Liu Estimator (RLE) by combining exact prior information with the sample information, and studied its properties. In the presence of stochastic restrictions, Theil and Goldberger (1961) proposed the Mixed Estimator (ME). By replacing OLSE by ME in the RE and LE respectively, the Stochastic Mixed Ridge Estimator (SMRE) (Li & Yang, 2010), and Stochastic Restricted Liu Estimator (SRLE) (Hubert & Wijekoon, 2006) were introduced.

When different estimators are available, preliminary test estimation procedure is adopted to select a suitable estimator. The preliminary test approach was first proposed by Bancroft (1944), and then has been studied by many researchers, such as Judge and Bock (1978), Wijekoon (1990), and Saleh and Kibria (1993). Later Golam Kibria and Saleh (2003) have discussed the performance of preliminary test ridge estimators based on large sample tests; WA (Wald, 1943), LR (Atchison & Silvey, 1958), and LM (Rao, 1947). Recently Arumairajan and Wijekoon (2013) proposed the Preliminary Test Stochastic Restricted Liu Estimator (PTSRLE) by combining Liu Estimator and Stochastic Restricted Liu Estimator. A Preliminary Test Principal Component Regression Estimator (PTPCRE) is proposed by combining the idea of Preliminary Test Estimator and Principal Component Regression Estimator.

Model Specification and Estimation

Consider the multiple linear model

$$y = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I) \quad (1)$$

where y is an $n \times 1$ observable random vector, X is an $n \times p$ known design matrix of rank p , β is a $p \times 1$ vector of unknown parameters and ε is an $n \times 1$ vector of disturbances.

The Ordinary Least Squares Estimator (OLSE) for the model (1) is given as

$$\hat{\beta} = S^{-1}X'y \quad (2)$$

where $S = X'X$.

Consider the transformation for model (1):

PRINCIPAL COMPONENT PRELIMINARY TEST ESTIMATOR

$$y = XTT'\beta + \varepsilon = Z\alpha + \varepsilon \quad (3)$$

where $Z = XT$, $\alpha = T'\beta$ and $T = (t_1, t_2, \dots, t_p) = (T_k, T_{p-k})$ is a $p \times p$ orthogonal matrix such that

$$(T_k, T_{p-k})' X'X (T_k, T_{p-k}) = \Lambda = \begin{pmatrix} \Lambda_k & 0 \\ 0 & \Lambda_{p-k} \end{pmatrix}$$

where $0 < k \leq p$, $T_k = (t_1, t_2, \dots, t_k)$, $T_{p-k} = (t_{k+1}, t_{k+2}, \dots, t_p)$, $\Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$, $\Lambda_k = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$, $\Lambda_{p-k} = \text{diag}(\lambda_{k+1}, \lambda_{k+2}, \dots, \lambda_p)$, and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$ are the eigenvalues of $X'X$. Note that $Z = XT = (z_1, z_2, \dots, z_p) = (Z_k, Z_{p-k})$ is the $n \times p$ matrix of the principal components, where $z_i = Xt_i$ is the i^{th} principal component. When Z_{p-k} contains principal components corresponding to near zero eigenvalues, Z can be separated as Z_k and Z_{p-k} , where Z_{p-k} is to be deleted. Rewrite the model (3) as

$$y = XTT'\beta = XT_k T_k' \beta + XT_{p-k} T_{p-k}' \beta + \varepsilon = Z_k \alpha_k + Z_{p-k} \alpha_{p-k} + \varepsilon. \quad (4)$$

By omitting Z_{p-k} , the OLSE of α_k is obtained, and $\hat{\alpha}_k = (Z_k' Z_k)^{-1} Z_k' y$. Then PCRE of β is

$$\hat{\beta}_{PCRE} = T_k (T_k' S T_k)^{-1} T_k' X' y \quad (5)$$

Xu and Yang (2011) showed that the PCRE estimator could be rewritten as follows.

$$\hat{\beta}_{PCRE} = T_k T_k' \hat{\beta} = L_k \hat{\beta} \quad (6)$$

where $L_k = T_k T_k'$.

The RE was proposed by Hoerl and Kennard (1970) as

$$\hat{\beta}(k) = W \hat{\beta} \quad (7)$$

where $W = (I + kS^{-1})^{-1}$ for $k \geq 0$.

The LE was introduced by Liu (1993) as

$$\hat{\beta}(d) = F_d \hat{\beta} \quad (8)$$

where $F_d = (S + I)^{-1} (S + dI)$ for $0 < d < 1$.

In addition to sample model (1), suppose some prior information was given about β in the form of a set of m independent stochastic linear restrictions as follows;

$$r = R\beta + \delta + v, \quad v \sim N(0, \sigma^2 \Omega) \quad (9)$$

where r is an $m \times 1$ stochastic known vector, R is a $m \times p$ of full row rank $m \leq p$ with known elements, δ is non zero $m \times 1$ unknown vector, v is an $m \times 1$ random vector of disturbances, and Ω is assumed to be known and positive definite. Further it is assumed that v is stochastically independent of ε i.e. $E(\varepsilon v') = 0$.

The Ordinary Least Squares Estimator (OLSE) for the model (1) and the Mixed Estimator (ME) (Theil & Goldberger, 1961) due to a stochastic prior restriction (9) are given by

$$\hat{\beta} = S^{-1} X' Y \text{ and } \hat{\beta}_m = \hat{\beta} + S^{-1} R' (\Omega + RS^{-1} R')^{-1} (r - R\hat{\beta}) \quad (10)$$

respectively. The expectation vector, and the mean square error matrix of $\hat{\beta}$ are given as $E(\hat{\beta}) = \beta$ and $MSE(\hat{\beta}) = \sigma^2 S^{-1}$ respectively.

The expectation vector, dispersion matrix, and the mean square error matrix of $\hat{\beta}_m$ are given as $E(\hat{\beta}_m) = \beta + H\delta$, $D(\hat{\beta}_m) = \sigma^2 S^{-1} - \sigma^2 G$ and $MSE(\hat{\beta}_m) = \sigma^2 (S^{-1} - G) + H\delta\delta'H'$ respectively, where, $G = S^{-1} R' (\Omega + RS^{-1} R')^{-1} RS^{-1}$, $H = S^{-1} R' (\Omega + RS^{-1} R')^{-1}$ and $\delta = E(r) - R\beta$.

Li and Yang (2010) proposed the Stochastic Mixed Ridge Estimator (SMRE), and is given as

$$\hat{\beta}_{SMRE}(k) = W \hat{\beta}_m. \quad (11)$$

PRINCIPAL COMPONENT PRELIMINARY TEST ESTIMATOR

The Stochastic Restricted Liu Estimator (SRLE) was proposed by Hubert and Wijekoon (2006), and is given by

$$\hat{\beta}_{srd}(d) = F_d \hat{\beta}_m. \quad (12)$$

By using the similar idea used by Hubert and Wijekoon (2006) and Li and Yang (2010), write the Stochastic Restricted Principal Component Regression Estimator (SRPCRE) as

$$\hat{\beta}_{SRPCRE} = L_k \hat{\beta}_m. \quad (13)$$

Turn to the question of the statistical evaluation of the compatibility of sample and stochastic information. The classical procedures is to test the hypothesis

$$H_0 : \delta = 0 \text{ against } H_1 : \delta \neq 0 \quad (14)$$

under linear model (1) and stochastic prior information (9).

The Ordinary Stochastic Pre Test Estimator (OSPE) of β (Wijekoon, 1990) is defined as

$$\hat{\beta}_{OSPE} = \begin{cases} \hat{\beta}_m & \text{if } H_0 : \delta = 0 \\ \hat{\beta} & \text{if } H_1 : \delta \neq 0 \end{cases} \quad (15)$$

Further, we can write equation (15) as

$$\hat{\beta}_{OSPE} = \hat{\beta}_m I_{[0, F_{m,n-p}(\alpha)]}(F) + \hat{\beta} I_{[F_{m,n-p}(\alpha), \infty)}(F) \quad (16)$$

$$\text{where } F = \frac{(r - R\hat{\beta})' (\Omega + RS^{-1}R')^{-1} (r - R\hat{\beta})}{m\hat{\sigma}^2} \quad (17)$$

has a non-central $F_{m,n-p,\lambda}$ distribution under $H_1 : \delta \neq 0$, with non-centrality parameter

$$\lambda = \frac{\delta'(\Omega + RS^{-1}R')^{-1}\delta}{2\sigma^2} \quad \text{with} \quad \hat{\sigma}^2 = \frac{(Y - X\hat{\beta})'(Y - X\hat{\beta})}{n-p} \quad (18)$$

and $I_{[0, F_{m,n-p}(\alpha)]}(F)$ and $I_{[0, F_{m,n-p}(\alpha), \infty)}(F)$ are indicator functions which take the value one if F falls in the subscripted interval, and zero otherwise. $F_{m,n-p}(\alpha)$ is the upper α -level critical value from the central F distribution $F_{m,n-p,0}$.

The expectation vector, dispersion matrix, and the mean square error matrix of $\hat{\beta}_{OSPE}$ are derived by Wijekoon (1990) are given below:

$$E(\hat{\beta}_{OSPE}) = \beta + h_{\lambda}(2)H\delta \quad (19)$$

$$D(\hat{\beta}_{OSPE}) = \sigma^2 S^{-1} - \sigma^2 h_{\lambda}(2)G + [2h_{\lambda}(2) - h_{\lambda}(4) - h_{\lambda}^2(2)]H\delta\delta'H' \quad (20)$$

and

$$MSE(\hat{\beta}_{OSPE}) = \sigma^2 S^{-1} - \sigma^2 h_{\lambda}(2)G + [2h_{\lambda}(2) - h_{\lambda}(4)]H\delta\delta'H' \quad (21)$$

respectively.

where $h_{\lambda}(\ell) = \Pr\left(\frac{\chi_{m+\ell, \lambda}^2}{\chi_{n-p}^2} \leq \frac{mF_{m,n-p}(\alpha)}{n-p}\right)$ for $\ell \in \mathbb{N}$.

When different estimators are available for the same parameter vector β in the linear regression model, one must solve the problem of their comparison. Usually as a simultaneous measure of covariance and bias, the mean square error matrix is used, and is defined by

$$MSE(\hat{\beta}, \beta) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] = D(\hat{\beta}) + B(\hat{\beta})B'(\hat{\beta}), \quad (22)$$

where $D(\hat{\beta})$ is the dispersion matrix and $B(\hat{\beta}) = E(\hat{\beta}) - \beta$ denotes the bias vector. Recall the Scalar Mean Square Error $SMSE(\beta, \hat{\beta}) = \text{trace}(MSE(\beta, \hat{\beta}))$.

PRINCIPAL COMPONENT PRELIMINARY TEST ESTIMATOR

For any two given estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, the estimator $\hat{\beta}_2$ is said to be superior to $\hat{\beta}_1$ under the MSEM criterion if and only if

$$M(\hat{\beta}_1, \hat{\beta}_2) = MSE(\hat{\beta}_1, \beta) - MSE(\hat{\beta}_2, \beta) \geq 0. \quad (23)$$

The Proposed Estimator

Now it is possible to propose the Preliminary Test Principal Component Regression Estimator (PTPCRE) as

$$\tilde{\beta}_{PTPCRE} = \begin{cases} L_k \hat{\beta}_m & \text{if } H_0 : \delta = 0 \\ L_k \hat{\beta} & \text{if } H_1 : \delta \neq 0 \end{cases} \quad (24)$$

Then the PTPCRE can be rewritten as follows.

$$\tilde{\beta}_{PTPCRE} = L_k \hat{\beta}_m I_{[0, F_{m, n-p}(\alpha))}(F) + L_k \hat{\beta} I_{(F_{m, n-p}(\alpha), \infty]}(F) = L_k \hat{\beta}_{OSPE}. \quad (25)$$

When $k = p$, L_k becomes I_p and consequently $\tilde{\beta}_{PTPCRE}$ becomes $\tilde{\beta}_{OSPE}$.

Using equations given in (19) and (20), we can now obtain the expectation vector, bias vector, dispersion matrix and mean square error matrix as

$$E(\tilde{\beta}_{PTPCRE}) = L_k \beta + h_\lambda(2) L_k H \delta \quad (26)$$

$$B(\tilde{\beta}_{PTPCRE}) = (L_k - I) \beta + h_\lambda(2) L_k H \delta \quad (27)$$

$$D(\tilde{\beta}_{PTPCRE}) = \sigma^2 L_k S^{-1} L_k' - \sigma^2 h_\lambda(2) L_k G L_k' + \left[2h_\lambda(2) - h_\lambda(4) - (h_\lambda(2))^2 \right] L_k H \delta \delta' H' L_k' \quad (28)$$

and

$$\begin{aligned}
MSE\left(\tilde{\beta}_{PTPCRE}\right) &= \sigma^2 L_k S^{-1} L_k' - \sigma^2 h_\lambda(2) L_k G L_k' \\
&+ \left[2h_\lambda(2) - h_\lambda(4) - (h_\lambda(2))^2 \right] L_k H \delta \delta' H' L_k' \\
&+ \left[(L_k - I) \beta + h_\lambda(2) L_k H \delta \right] \left[(L_k - I) \beta + h_\lambda(2) L_k H \delta \right]'
\end{aligned} \tag{29}$$

respectively.

The Proposed Estimator Based on WA, LR and LM Tests

In general, the finite sample tests such as t or F were used to define the preliminary test estimator. In the field of Econometrics, these finite sample tests are not used due to large samples. In this situation it is very useful to define preliminary test estimators based on large sample tests. The three large sample tests considered in the literature are WA, LR, and LM. The WA test offers the advantage of only requiring estimates of the unrestricted model, whereas LR test requires estimates of both unrestricted and the restricted model. The LM test only requires estimates of the restricted model. In different situations, we may use one of these tests which are easier to compute.

Judge and Bock (1978) have rewritten the model given in (1) and (9) to obtain the F statistics for testing the hypothesis in (14). Using the rewritten model we can derive the test statistics for the WA, the LR and the LM tests which are well employed for testing the hypothesis (14), and are given by

$$\varepsilon_{WA} = \frac{(n+m)mF}{n-p}, \quad \varepsilon_{LR} = (n+m) \ln \left[1 + \frac{mF}{n-p} \right], \quad \text{and} \quad \varepsilon_{LM} = \frac{(n+m)mF}{n-p+mF} \tag{30}$$

respectively (Evans & Savin, 1982).

It's known that under the null hypothesis H_0 , the three test statistics have the same asymptotic chi-square distribution with m degrees of freedom (Evans & Savin, 1982). When the exact distribution is approximated by the asymptotic chi-square distribution, the critical value for an α level test of H_0 is approximated by the central chi-square critical value $\chi_m^2(\alpha)$ for large sample tests. Further Berndt and Savin (1977) showed that a symmetric numerical inequality $\varepsilon_{WA} \geq \varepsilon_{LR} \geq \varepsilon_{LM}$ exists between these three tests. This asymptotic chi-square distribution has wide applications in the field of Econometrics.

Based on the above tests, the PTPCRE takes the form

$$\tilde{\beta}_{PTPCRE}(\varepsilon_*) = L_k \hat{\beta}_m I_{[0, \chi_m^2(\alpha)]}(\varepsilon_*) + L_k \hat{\beta} I_{[\chi_m^2(\alpha), \infty)}(\varepsilon_*) \quad (31)$$

where (*) stands for either WA, LR or LM tests values, and $\chi_m^2(\alpha)$ is the upper percentiles of the central χ^2 distribution with m degrees of freedom.

Using equation (26), (27), (28) and (29) we can obtain the stochastic properties of PTPCRE based on WA, LR and LM tests as follows.

$$E[\tilde{\beta}_{PTPCRE}(\varepsilon_*)] = L_k \beta + h_\lambda^*(2) L_k H \delta, \quad (32)$$

$$B[\tilde{\beta}_{PTPCRE}(\varepsilon_*)] = (L_k - I) \beta + h_\lambda^*(2) L_k H \delta, \quad (33)$$

$$\begin{aligned} D[\tilde{\beta}_{PTPCRE}(\varepsilon_*)] &= \sigma^2 L_k S^{-1} L_k' - \sigma^2 h_\lambda^*(2) L_k G L_k' \\ &\quad + \left[2h_\lambda^*(2) - h_\lambda^*(4) - (h_\lambda^*(2))^2 \right] L_k H \delta \delta' H' L_k', \end{aligned} \quad (34)$$

$$\begin{aligned} MSE[\tilde{\beta}_{PTPCRE}(\varepsilon_*)] &= \sigma^2 L_k S^{-1} L_k' - \sigma^2 h_\lambda^*(2) L_k G L_k' \\ &\quad + \left[2h_\lambda^*(2) - h_\lambda^*(4) - (h_\lambda^*(2))^2 \right] L_k H \delta \delta' H' L_k' \\ &\quad + \left[(L_k - I) \beta + h_\lambda^*(2) L_k H \delta \right] \left[(L_k - I) \beta + h_\lambda^*(2) L_k H \delta \right]', \end{aligned} \quad (35)$$

where $h_\lambda^*(\ell) = \Pr\left(\frac{\chi_{m+\ell, \lambda}^2}{\chi_{n-p}^2} \leq \frac{mc^*}{n-p}\right)$ for $\ell \in N$ and c^* takes the value for WA, LR and LM tests as

$$c^{WA} = \frac{(n-p)\chi_m^2(\alpha)}{(n+m)m}, \quad c^{LR} = \frac{(n-p)\left(e^{\chi_m^2(\alpha)/(n+m)} - 1\right)}{m} \quad \text{and}$$

$$c^{LM} = \frac{(n-p)\chi_m^2(\alpha)}{m(n+m-\chi_m^2(\alpha))}.$$

Mean Square Error Matrix Comparisons

The performance of PTPCRE will be compared using WA, LR and LM tests with respect to Mean Square Error Matrix (MSEM) sense for the two cases in which the stochastic restrictions are correct, and not correct.

Now we consider the following dispersion matrix differences.

$$D[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})] - D[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})] = \sigma^2 \psi_1 L_k G L'_k - \xi_1 L_k H \delta \delta' H' L'_k,$$

$$D[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})] - D[\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})] = \sigma^2 \psi_2 L_k G L'_k - \xi_2 L_k H \delta \delta' H' L'_k,$$

$$D[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})] - D[\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})] = \sigma^2 \psi_3 L_k G L'_k - \xi_3 L_k H \delta \delta' H' L'_k,$$

where

$$\psi_1 = h_{\lambda}^{LR}(2) - h_{\lambda}^{WA}(2) \geq 0, \quad \psi_1^* = h_{\lambda}^{LR}(4) - h_{\lambda}^{WA}(4) \geq 0, \quad \tilde{\psi}_1 = h_{\lambda}^{LR}(2) + h_{\lambda}^{WA}(2) \geq 0,$$

$$\psi_2 = h_{\lambda}^{LM}(2) - h_{\lambda}^{LR}(2) \geq 0, \quad \psi_2^* = h_{\lambda}^{LM}(4) - h_{\lambda}^{LR}(4) \geq 0,$$

$$\tilde{\psi}_2 = h_{\lambda}^{LM}(2) + h_{\lambda}^{LR}(2) \geq 0, \quad \psi_3 = h_{\lambda}^{LM}(2) - h_{\lambda}^{WA}(2) \geq 0,$$

$$\psi_3^* = h_{\lambda}^{LM}(4) - h_{\lambda}^{WA}(4) \geq 0, \quad \tilde{\psi}_3 = h_{\lambda}^{LM}(2) + h_{\lambda}^{WA}(2) \geq 0, \quad \xi_1 = (2 - \psi_1) \tilde{\psi}_1 - \psi_1^*,$$

$$\xi_2 = (2 - \psi_2) \tilde{\psi}_2 - \psi_2^* \quad \text{and} \quad \xi_3 = (2 - \psi_3) \tilde{\psi}_3 - \psi_3^*.$$

It is clear that $2 - \psi_1 \geq 1$ since $\psi_1 \leq 1$. This implies that $(2 - \psi_1) \tilde{\psi}_1 \geq \tilde{\psi}_1$ as $\tilde{\psi}_1 \geq 0$. But we can show that $\tilde{\psi}_1 - \psi_1^* \geq 0$. This implies that $\xi_1 = (2 - \psi_1) \tilde{\psi}_1 - \psi_1^* \geq 0$ as $(2 - \psi_1) \tilde{\psi}_1 \geq \tilde{\psi}_1$. Similarly we can show that $\xi_2 \geq 0$ and $\xi_3 \geq 0$.

Write the following mean square error matrices differences.

$$\begin{aligned} &MSE[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})] - MSE[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})] \\ &= L_k (D_1 + b_{WA} b'_{WA} - b_{LR} b'_{LR}) L'_k, \end{aligned} \tag{36}$$

$$\begin{aligned} &MSE\left[\tilde{\beta}_{PTPCRE}\left(\varepsilon_{LR}\right)\right]-MSE\left[\tilde{\beta}_{PTPCRE}\left(\varepsilon_{LM}\right)\right] \\ &=L_k\left(D_2+b_{LR}b'_{LR}-b_{LM}b'_{LM}\right)L'_k, \end{aligned} \quad (37)$$

$$\begin{aligned} &MSE\left[\tilde{\beta}_{PTPCRE}\left(\varepsilon_{WA}\right)\right]-MSE\left[\tilde{\beta}_{PTPCRE}\left(\varepsilon_{LM}\right)\right] \\ &=L_k\left(D_3+b_{WA}b'_{WA}-b_{LM}b'_{LM}\right)L'_k, \end{aligned} \quad (38)$$

where $D_i = \sigma^2 \psi_i G - \xi_i H \delta \delta' H'$ for $i = 1, 2, 3$, $b_{WA} = (I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta$, $b_{LR} = (I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta$ and $b_{LM} = (I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta$.

Note that $b_{WA} = b_{LR} = b_{LM} = (I - L_k^{-1})\beta$ when $\delta = 0$.

Based on the mean square error matrix differences the following theorems can be stated.

Theorem 1:

- i) When the stochastic restrictions are true (i.e. $\delta = 0$), $\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$ is always superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$ in the mean square error matrix sense.
- ii) When the stochastic restriction are not true (i.e. $\delta \neq 0$), and if $\lambda \leq \frac{\psi_1}{2[(2-\psi_1)\tilde{\psi}_1 - \psi_1^*]}$ then the $\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$ is superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$ with respect to MSE matrix sense if and only if

$$\begin{aligned} &\left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right] + 1 \right\} \\ &* \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right] - 1 \right\} \\ &\leq \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right] \right\}^2. \end{aligned}$$

Proof:

Consider the mean square error matrix difference (36) between WA and LR.

$$MSE\left[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})\right] - MSE\left[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})\right] = L_k (D_1 + b_{WA}b'_{WA} - b_{LR}b'_{LR})L'_k$$

where $D_1 = \sigma^2\psi_1 G - \xi_1 H\delta\delta'H'$.

When the stochastic restrictions are true (i.e. $\delta = 0$), then the MSE matrix difference in (36) reduces to $\sigma^2\psi_1 L_k GL'_k$ which is clearly a nonnegative definite matrix since $\psi_1 \geq 0$, $G \geq 0$ and $L_k > 0$.

When the stochastic restrictions are not correct (i.e. $\delta \neq 0$), then $MSE\left[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})\right] - MSE\left[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})\right]$ is a nonnegative definite matrix if and only if $D_1 + b_{WA}b'_{WA} - b_{LR}b'_{LR}$ is a nonnegative matrix. To apply lemma 3 (Appendix) we have to show that D_1 is a nonnegative definite matrix. We rewrite D_1 as $D_1 = \xi_1 D_1^*$, where $D_1^* = \frac{\sigma^2\psi_1}{\xi_1} G - H\delta\delta'H'$. Then D_1 is a nonnegative definite matrix, if and only if D_1^* is nonnegative definite matrix. To show that D_1^* is nonnegative definite matrix, lemma 1 (Appendix) is used by setting

$$\gamma = \frac{\sigma^2\psi_1}{\xi_1}, \quad B = G \quad \text{and} \quad a = H\delta.$$

Note that $G = S^{-1}R'(\Omega + RS^{-1}R')^{-1}RS^{-1} \geq 0$, and the generalized inverse of G is $G^- = SR^+(\Omega + RS^{-1}R')^{-1}R'^+S$. Hence $GG^-H\delta = H\delta$. This implies that $H\delta \in \mathfrak{R}(G)$, where $\mathfrak{R}(\cdot)$ denote the column space of the corresponding matrix and R^+ is a Moore-Penrose matrix of R .

Then according to lemma 1, $D_1^* \geq 0$ if and only if

$$\delta'H'G^-H\delta \leq \frac{\sigma^2\psi_1}{\xi_1}. \quad (39)$$

After some straightforward calculations it can be shown

$$\delta' H' G^- H \delta = \delta' (\Omega + R S^{-1} R')^{-1} \delta. \quad (40)$$

By substituting this result to (39) obtain

$$\frac{\delta' (\Omega + R S^{-1} R')^{-1} \delta}{2\sigma^2} \leq \frac{\psi_1}{2\xi_1}. \quad (41)$$

Using (18), this inequality can be rewritten as

$$\lambda \leq \frac{\psi_1}{2\xi_1} = \frac{\psi_1}{2[(2 - \psi_1)\tilde{\psi}_1 - \psi_1^*]}. \quad (42)$$

This implies that D_1^* is a nonnegative definite matrix if and only if

$$\lambda \leq \frac{\psi_1}{2[(2 - \psi_1)\tilde{\psi}_1 - \psi_1^*]}. \text{ Therefore } D_1 = \xi_1 L_k D_1^* L_k' \text{ is nonnegative definite matrix}$$

$$\text{if and only if } \lambda \leq \frac{\psi_1}{2[(2 - \psi_1)\tilde{\psi}_1 - \psi_1^*]}, \xi_1 \geq 0.$$

To apply lemma 3 (Appendix), the Moore Penrose inverse of D_1 is obtained by using lemma 2 (Appendix), and is given by

$$D_1^+ = \frac{1}{\sigma^2 \psi_1} \times \left[G^+ + \frac{\xi_1}{\sigma^2 \psi_1 - \xi_1 \delta' H' G^+ H \delta} G^+ H \delta \delta' H' G^+ \right] \quad (43)$$

After some straightforward calculations we can show that

$$\delta' H' G^+ H \delta = 2\sigma^2 \lambda \quad (44)$$

Using (43) and (44) we can easily prove that $D_1 D_1^+ = I_p$, where I_p is an identity matrix with order $(p \times p)$. This implies that $D_1 D_1^+ b_{WA} = b_{WA}$ and $D_1 D_1^+ b_{LR} = b_{LR}$. Then we have $b_{WA} \in \mathfrak{R}(D_1)$ and $b_{LR} \in \mathfrak{R}(D_1)$. To establish condition (a) in lemma 3, we find $f_{ij} = b_i' D_1^- b_j$ for $i, j = WA, LR$ such that

$$f_{WA,WA} = \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right]$$

$$f_{LR,LR} = \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right] \text{ and}$$

$$f_{WA,LR} = \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right]$$

Note that, instead of D_1^- , the Moore Penrose inverse D_1^+ of D_1 is used, since f_{ij} is invariant to the choice of D_1^- .

Now according to [lemma 3 \(Appendix\)](#)

$$MSE[\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})] - MSE[\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})] \geq 0 \text{ if and only if}$$

$$\begin{aligned} & \left\{ \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right] + 1 \right\} \\ & * \left\{ \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right] - 1 \right\} \\ & \leq \left\{ \left[(I - L_k^{-1})\beta + h_\lambda^{WA}(2)H\delta \right]' D_1^+ \left[(I - L_k^{-1})\beta + h_\lambda^{LR}(2)H\delta \right] \right\}^2. \end{aligned}$$

This completes the proof.

By considering the mean square error matrix differences given in equation (37) and (38), we can state [Theorem 2](#) and [Theorem 3](#) respectively. The proofs of these theorems are similar to the proof of [Theorem 1](#).

Theorem 2:

- i) When the stochastic restrictions are true (i.e. $\delta = 0$), $\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$ is always superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$ in the mean square error matrix sense.

- ii) When the stochastic restriction are not true (i.e. $\delta \neq 0$), and if $\lambda \leq \frac{\psi_2}{2[(2-\psi_2)\tilde{\psi}_2 - \psi_2^*]}$ then the $\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$ is superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$ with respect to MSE matrix sense if and only if

$$\begin{aligned} & \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right]' D_2^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right] + 1 \right\} \\ & * \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right]' D_2^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right] - 1 \right\} \\ & \leq \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LR}(2)H\delta \right]' D_2^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right] \right\}^2. \end{aligned}$$

Theorem 3:

- i) When the stochastic restrictions are true (i.e. $\delta = 0$), $\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$ is always superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$ in the mean square error matrix sense.
- ii) When the stochastic restrictions are not true (i.e. $\delta \neq 0$), and if $\lambda \leq \frac{\psi_3}{2[(2-\psi_3)\tilde{\psi}_3 - \psi_3^*]}$ then the $\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$ is superior to $\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$ with respect to MSE matrix sense if and only if

$$\begin{aligned} & \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right]' D_3^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right] + 1 \right\} \\ & * \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right]' D_3^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right] - 1 \right\} \\ & \leq \left\{ \left[(I - L_k^{-1})\beta + h_{\lambda}^{WA}(2)H\delta \right]' D_3^+ \left[(I - L_k^{-1})\beta + h_{\lambda}^{LM}(2)H\delta \right] \right\}^2. \end{aligned}$$

Monte Carlo Simulation

To illustrate the behavior of the proposed estimators, a Monte Carlo Simulation study was designed by considering different levels of multicollinearity. Following McDonald and Galarneau (1975) generate explanatory variables as follows:

$$x_{ij} = (1 - \rho^2)^{1/2} z_{ij} + \rho z_{i,p+1}, \quad i = 1, 2, \dots, n, \quad j = 1, 2, \dots, p,$$

where z_{ij} is an independent standard normal pseudo random number, and ρ is specified so that the theoretical correlation between any two explanatory variables is given by ρ^2 . A dependent variable is generated by using the equation.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i, \quad i = 1, 2, \dots, n,$$

where ε_i is a normal pseudo random number with mean zero and variance σ_i^2 . Newhouse and Oman (1971) have noted that if the MSE is a function of σ^2 and β , and if the explanatory variables are fixed, then subject to the constraint $\beta' \beta = 1$, the MSE is minimized when β is the normalized eigenvector corresponding to the largest eigenvalue of the $X'X$ matrix. In this study we choose the normalized eigenvector corresponding to the largest eigenvalue of $X'X$ as the coefficient vector β , $n = 50$, $p = 4$ and $\sigma_i^2 = 1$. Three different sets of correlations are considered by selecting the values as $\rho = 0.7, 0.8$ and 0.9 , and two various significance levels are taken as $\alpha = 0.01$ and 0.05 . Further R , r and v in equation (9) are taken as $R = (0, 1, 3, 1)$, $r = 0$ and $v \sim N(0, \Omega = \hat{\sigma}_\rho^2)$, where $\hat{\sigma}_\rho^2$ is estimated by using equation (18). The eigenvalues of the matrix S for $\rho = 0.7, 0.8$ and 0.9 are given in Table 1.

The first three principal components account for 91.29% and 93.65% of the total variance when $\rho = 0.7$ and 0.8 respectively, and also the first two principal components account for 91.8% of the total variance when $\rho = 0.9$. Therefore we choose the number of the principal components $k = 3$ when $\rho = 0.7$ and 0.8 , and $k = 2$ when $\rho = 0.9$. Table 2, Table 3 and Table 4 show the scalar mean square errors (SMSE) obtained by using equation (35).

PRINCIPAL COMPONENT PRELIMINARY TEST ESTIMATOR

Table 1. Eigenvalues of the matrix S for $\rho = 0.7, 0.8$ and 0.9 .

ρ	Eigenvalues	Proportion of Total Variance (%)	Cumulative Percentage of Total Variance (%)
0.7	0.12261	63.02	63.02
	0.03209	16.49	79.51
	0.02292	11.78	91.29
	0.01693	8.80	100
0.8	0.13843	73.15	73.15
	0.02255	11.92	85.07
	0.01624	8.58	93.65
	0.01201	6.35	100
0.9	0.15549	85.30	85.3
	0.01184	6.50	91.8
	0.00861	4.72	96.52
	0.00636	3.48	100

Table 2. Estimated SMSE of PTPCRE for WA, LR and LM tests for $\rho = 0.7$ and $k = 3$.

Estimators	SMSE at $\alpha = 0.01$	SMSE at $\alpha = 0.05$
$\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$	0.0584	0.0543
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$	0.0590	0.0544
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$	0.0597	0.0544

Table 3. Estimated SMSE of PTPCRE for WA, LR and LM tests for $\rho = 0.8$ and $k = 3$.

Estimators	SMSE at $\alpha = 0.01$	SMSE at $\alpha = 0.05$
$\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$	0.2444	0.1556
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$	0.2621	0.1592
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$	0.2841	0.1633

Table 4. Estimated SMSE of PTPCRE for WA, LR and LM tests for $\rho = 0.9$ and $k = 2$.

Estimators	SMSE at $\alpha = 0.01$	SMSE at $\alpha = 0.05$
$\tilde{\beta}_{PTPCRE}(\varepsilon_{WA})$	0.3227	1.7228
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LR})$	0.3313	1.7098
$\tilde{\beta}_{PTPCRE}(\varepsilon_{LM})$	0.3409	1.6961

Based on Table 2, there are no big differences in the SMSE among the estimators when $\rho = 0.7$. Based on the Table 3, the PTPCRE based on WA test has the smallest SMSE. From Table 4, notice that when $\rho = 0.9$ and $\alpha = 0.01$, the PTPCRE based on WA test has the smallest SMSE. Then the LM test has the smallest SMSE when $\rho = 0.9$ and $\alpha = 0.05$.

Conclusion

A new Preliminary Test Estimator based on Principal Component Regression Estimator defined in the linear regression model when the stochastic restrictions are available in addition to the sample information, and when the explanatory variables are multicollinear. Based on the simulation study, we can conclude that the PTPCRE based on WA test has the smallest SMSE when $\rho = 0.9$ and $\alpha = 0.01$. The PTPCRE based on LM test has smallest SMSE when $\rho = 0.9$ and $\alpha = 0.05$.

References

- Arumairajan, S & Wijekoon, P. (2013). Improvement of the preliminary test estimator when stochastic restrictions are available in linear regression model. *Open Journal of Statistics*, 3(4), 283-292. doi:[10.4236/ojs.2013.34033](https://doi.org/10.4236/ojs.2013.34033)
- Atchison, J & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics*, 29(3), 813-828. doi:[10.1214/aoms/1177706538](https://doi.org/10.1214/aoms/1177706538)
- Baksalary, J. K. & Kala, R. (1983). Partial orderings between matrices one of which is of rank one. *Bulletin of the Polish Academy of Science, Mathematics*, 31, 5-7.
- Baksalary, J. K. & Trenkler, G. (1991). Nonnegative and positive definiteness of matrices modified by two matrices of rank one. *Linear Algebra and its Applications*, 151, 169-184. doi:[10.1016/0024-3795\(91\)90362-Z](https://doi.org/10.1016/0024-3795(91)90362-Z)
- Bancroft, T. A. (1944). On biases in estimation due to use of preliminary tests of significance. *Annals of Mathematical Statistics*, 15(2), 190-204. doi:[10.1214/aoms/1177731284](https://doi.org/10.1214/aoms/1177731284)
- Berndt, E. R. & Savin, N. E. (1977). Conflict among criteria for testing hypothesis in the multivariate linear regression model. *Econometrica*, 45(5), 1263-1278. doi:[10.2307/1914072](https://doi.org/10.2307/1914072)

PRINCIPAL COMPONENT PRELIMINARY TEST ESTIMATOR

- Evans, G. B. A. & Savin, N. E. (1982). Conflict among the criteria revisited; The W, LR and LM tests. *Econometrica*, 50(3), 737-748. doi:[10.2307/1912611](https://doi.org/10.2307/1912611)
- Golam Kibria, B. M. & Saleh, A. K. Md. E. (2003) Effect of W, LR and LM tests on the performance of preliminary test ridge regression estimators. *Journal of the Japan Statistical Society*, 33 (1), 119-136. doi:[10.14490/jjss.33.119](https://doi.org/10.14490/jjss.33.119)
- Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. doi:[10.2307/1267351](https://doi.org/10.2307/1267351)
- Hubert, M. H. & Wijekoon, P. (2006). Improvement of the Liu estimator in linear regression model. *Statistical Papers*, 47(3), 471-479. doi:[10.1007/s00362-006-0300-4](https://doi.org/10.1007/s00362-006-0300-4)
- Judge, G. G. & Bock, M. E. (1978). *The statistical implications of pre-test and Stein-rule estimators in econometrics*. New York: North Holland.
- Kaçiranlar, S., Sakallioğlu, S., Akdeniz, F., Styan, G. P. H. & Werner, H. J. (1999). A new biased estimator in linear regression and a detailed analysis of the widely analyzed dataset on Portland cement. *The Indian Journal of Statistics*, 61(3), 443-459.
- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communication in Statistics*, 32(2), 419-435. doi:[10.1081/SAC-120017499](https://doi.org/10.1081/SAC-120017499)
- Li, Y. & Yang, H. (2010). A new stochastic mixed ridge estimator in linear regression. *Statistical Papers*, 51(2), 315-323. doi:[10.1007/s00362-008-0169-5](https://doi.org/10.1007/s00362-008-0169-5)
- Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, 22(2), 393-402. doi:[10.1080/03610929308831027](https://doi.org/10.1080/03610929308831027)
- Massy, W. F. (1965). Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309), 234-266. doi:[10.1080/01621459.1965.10480787](https://doi.org/10.1080/01621459.1965.10480787)
- McDonald, G. C. & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of American Statistical Association*, 70(350), 407-416. doi:[10.1080/01621459.1975.10479882](https://doi.org/10.1080/01621459.1975.10479882)
- Newhouse, J. P. & Oman, S. D. (1971). An evaluation of ridge estimators. *Rand Report, R-716-Pr*, 1-28.
- Rao, R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society*, 44(1), 50-57.

Saleh, A. K. M. E. & Kibria, B. M. G. (1993). Performance of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics—Theory and Methods*, 22(10), 2747-2764.

doi:[10.1080/03610929308831183](https://doi.org/10.1080/03610929308831183)

Sarkar, N. (1992). A new estimator combining the ridge regression and the restricted least squares methods of estimation. *Communication in Statistics-Theory Methods*, 21(7), 1987-2000. doi:[10.1080/03610929208830893](https://doi.org/10.1080/03610929208830893)

Theil, H. & Goldberger, A. S. (1961). On pure and mixed estimation in economics. *International Economic review*, 2(1), 65-78. doi:[10.2307/2525589](https://doi.org/10.2307/2525589)

Trenkler, G. (1985). Mean square error matrix comparisons of estimators in linear regression. *Communication in Statistics A*, 14(10), 2495-2509.

doi:[10.1080/03610928508829058](https://doi.org/10.1080/03610928508829058)

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transaction of the American Mathematical Society*, 54(3), 426-482. doi:[10.2307/1990256](https://doi.org/10.2307/1990256)

Wijekoon, P. (1990). *Mixed estimation and preliminary test estimation in the linear regression model*. (Unpublished doctoral thesis). University of Dortmund, Germany.

Xu, J. & Yang, H. (2011). On the restricted $r-k$ class estimator and the restricted $r-d$ class estimator in linear regression. *Journal of Statistical Computation and Simulation*, 81(6), 679-691. doi:[10.1080/00949650903471023](https://doi.org/10.1080/00949650903471023)

Appendix

Lemma 1: (Baksalary & Kala, 1983)

Suppose B is a symmetric real $(n \times n)$ matrix, a is an $(n \times 1)$ real vector and γ is a positive real number. Then the following two properties are equivalent

- i) $\gamma B - aa'$ is nonnegative definite (n.n.d)
- ii) B is n.n.d, $a \in \mathfrak{R}(B)$ and $a'B^-a \leq \gamma$.

Lemma 2: (Trenkler, 1985)

Let A be a symmetric $(n \times n)$ matrix, and let a , a_1 , and a_2 be $(n \times 1)$ vectors. Suppose that

- a) $a \in \mathfrak{R}(A)$, and the real numbers ϕ and ψ satisfy $\phi \neq 0$ and $\phi + \psi a'A^+a \neq 0$. Then we have the identity

$$[\phi A + \psi aa']^+ = \frac{1}{\phi} \left[A^+ - \frac{\psi}{\phi + \psi a'A^+a} A^+ aa'A^+ \right]$$
- b) $a_j \in \mathfrak{R}(A)$, $j = 1, 2$, and the real number ρ satisfies $1 + \rho a_1'A^+a_1 \neq 0$. Then we have $a_2 \in \mathfrak{R}(A + \rho a_1 a_1')$.

Lemma 3: (Baksalary & Trenkler, 1991)

Let C be a nonnegative definite matrix and c_1, c_2 be linearly independent vectors. Furthermore for some generalized inverse C^- of C , let $f_{ij} = c_i' C^- c_j$; $i = 1, 2, j = 1, 2$ and let

$$s = \frac{c_2'(I - CC^-)'(I - CC^-)c_2}{c_1'(I - CC^-)(I - CC^-)c_1}$$

where $c_1 \in \mathfrak{R}(C)$ and $\mathfrak{R}(\cdot)$ denote the column space of the corresponding matrix. Then we have $C + c_1 c_1' - c_2 c_2' \geq 0$ if and only if

- a) $c_1 \in \mathfrak{R}(C), c_2 \in \mathfrak{R}(C)$ and $(f_{11} + 1)(f_{22} - 1) \leq f_{12}^2$ or
- b) $c_1 \notin \mathfrak{R}(C), c_2 \in \mathfrak{R}(C, c_1)$ and $(c_2 - s c_1)C^-(c_2 - s c_1) \leq 1 - s^2$

and all expressions in (a) and (b) are independent of the choice of C^- .

Symmetric Variants of Logistic Smooth Transition Autoregressive Models: Monte Carlo Evidences

OlaOluwa S. Yaya

University of Ibadan
Ibadan, Nigeria

Olanrewaju I. Shittu

University of Ibadan
Ibadan, Nigeria

The Smooth Transition Autoregressive (STAR) models are becoming popular in modeling economic and financial time series. The asymmetric type of the model is the Logistic STAR (LSTAR) model, which is limited in its applications as a result of its asymmetric property, which makes it suitable for modelling specific macroeconomic time series. This study was designed to develop the Absolute Error LSTAR (AELSTAR) and Quadratic LSTAR (QLSTAR) models for improving symmetry and performance in terms of model fitness. Modified Teräsvirta's Procedure (TP) and Escibano and Jordá's Procedure (EJP) were used to test for nonlinearity in the series. The performance of the AELSTAR and QLSSTAR models showed that TP and EJP realized time series with improved symmetry as indicated by the lower relative frequencies than that realized with the existing LSTAR model. The AELSTAR model performed better than QLSSTAR model at higher nonlinearity, and the selection of both models showed evidence of asymptotic property. The AELSTAR and QLSSTAR models showed improved symmetry over the existing asymmetric LSTAR model.

Keywords: Nonlinear models, smooth transition autoregressive models, transition function

---

Introduction

Smooth Transition Autoregressive (STAR) models have found widespread application in economics and finance. Logistic STAR (LSTAR), a form of the model, is characterized by the asymmetric properties which make it suitable for modelling specific macroeconomic time series. The asymmetric property often limits its application to some symmetric time series. The study was designed to develop the Absolute Error LSTAR (AELSTAR) and Quadratic LSTAR (QLSTAR) models for improving symmetry and performance.

---

*Dr. Yaya is a Lecturer in the Department of Statistics. Email him at: os.yaya@ui.edu.ng.
Dr. Shittu is an Associate Professor of Statistics. Email him at: oi.shittu@ui.edu.ng.*

LOGISTIC SMOOTH TRANSITION

The STAR model was introduced into time series literature by Chan and Tong (1986), who used the cumulative function of the standard normal variable as the transition function in the model. The specification, estimation, and evaluation of the model were considered in Teräsvirta (1994). This model was used to describe non-linearity in the business cycle (Teräsvirta & Anderson, 1992; Skalin & Teräsvirta, 1996; 1998) and real exchange rates (Baum, Caglayan, & Barkoulas, 1998; Liew, Ahmad, & Sie-Hoe, 2002). Acemoglu and Scott (1994) examined the connection between business cycle, nonlinearity, and asymmetry in the UK labor market and found that the variables are interrelated. Öcal (2000) applied the STAR model on the nonlinearities in growth rates of UK macroeconomic time series: GDP, price, consumption, retail sales, personal disposable income, savings, investment, industrial production, and unemployment, and their findings suggest a three-regime STAR model for modelling GDP, price, and consumption. UK stock market returns have been studied using the Smooth Transition Regression (STR) framework by employing a variety of financial and macroeconomic series that are assumed to influence UK stock returns, namely GDP, interest rates, inflation, money supply, and US stock prices (Aslanidis, Osborn, & Sensier, 2002). They estimated STR models where the linearity hypothesis is strongly rejected for at least one transition variable. These non-linear models described the in-sample movements of the stock returns series better than the corresponding linear model.

More recent applications are Teräsvirta, van Dijk, and Medeiros (2005), Woodward and Anderson (2009), and Dueker, Sola, and Spagnolo (2006). Teräsvirta et al. (2005) examine the forecast accuracy of linear Autoregressive (AR) and STAR models and concluded that STAR model generally outperforms linear AR models. Dueker et al. (2006) worked on the STAR model and proposed the Contemporaneous Smooth Threshold Autoregressive (C-STAR), model which is a modification of Teräsvirta (1994). This C-STAR model does not require the initial regime to be predetermined and was successfully applied to interest rate modelling. The results indicated that the model is capable of outperforming some competing alternative nonlinear models, especially in terms of relative out of sample forecasting performance. Woodward and Anderson (2009) studied the behavior of the financial markets using the LSTAR model to classify the market in two phases of bull and bear and the movement in the bull and bear were not the same, which confirmed asymmetry in the markets.

The STAR model is of two forms: the Logistic STAR (LSTAR) and Exponential STAR (ESTAR), which have asymmetric and symmetric properties, respectively. Though market data are often asymmetric, the possibility of improving the symmetry of the LSTAR model could lead to models with better

parameter estimates and forecasts. Liew et al. (2002) proposed the Absolute Logistic STAR model with transition function that allows a V-shaped symmetry adjustment towards the mean of the series. In their model, mean adjusted data were used, and the performance of their model outperformed that of the LSTAR model. Adebile (2007) proposed the Error Logistic smooth Transition Regression (ELSTR) model with improved asymmetry over the existing LSTAR, and the model performed better in terms of forecasts.

The economic and financial data display different levels of asymmetry; the possibility of improving the asymmetry of LSTAR models may lead to improvement of the model in terms of parameter estimates and forecasts. The performance of the proposed AELSTAR and QLSTAR models in terms of symmetry are then judged based on their similarity in their realization to symmetric ESTAR model.

The remaining part of the work is structured as follows: the STAR model and the proposed models are discussed in following sections; the linearity and model specification testing procedures are then explained; the Monte Carlo simulations are presented; while a final section gives the conclusion.

The General STAR Model

Following van Dijk, Teräsvirta, and Franses (2002), the general two-regime STAR model of order p observed at $t = 1 - p, 1 - (p - 1), \dots, -1, 0, 1, N - 1, N$ for a univariate time series y_t is

$$y_t = \phi_1 \tilde{\mathbf{y}}_t^{(p)} (1 - F(s_t; \gamma, c)) + \phi_2 \tilde{\mathbf{y}}_t^{(p)} F(s_t; \gamma, c) + \epsilon_t, \quad (1)$$

where $\tilde{\mathbf{y}}_t^{(p)} = (1 - y_{t-1}, \dots, y_{t-p})'$, $\phi_i = (\phi_{i0}, \phi_{i1}, \dots, \phi_{ip})'$, and $i = \{1, 2\}$. The ϵ_t are assumed to be a difference sequence with respect to the history of the time series up to time $t - 1$, denoted by $\Omega_{t-1} = \{y_{t-1}, y_{t-2}, \dots, y_{1-(p-1)}, y_{t-p}\}$. That is $E(\epsilon_t^2 | \Omega_{t-1}) = \sigma^2$ and $E(\epsilon_t | \Omega_{t-1}) = 0$ or $E(\epsilon_t^2 | \Omega_{t-1}) = \sigma_t^2$ for the heteroscedastic STAR model.

Following Tsay (2005), the transition function $F(y_{t-d}; \gamma, c)$ can be a logistic, exponential, or cumulative continuous distribution function third order continuously differentiable with respect to γ (Escribano & Jordá, 1999; 2001), and is bounded between 0 and 1 for both LSTAR and ESTAR specifications. It defines regime-specific dynamics that govern the transition between 0 and 1 regimes, depending on the values of the transition variable relative to the slope γ and of the

location parameters c . The smoothness parameter $\gamma > 0$ determines the speed of transition of the transition function towards the inner or outer regime, as well as the degree of nonlinearity. As $\gamma \rightarrow 0$, both LSTAR and ESTAR functions converge to constant, and the models become linear AR models. The delay parameter d of the transition variable can take values in the range of $1 \leq d \leq p$ or $d > p$ (van Dijk, 1999; Siliverstovs, 2005).

The LSTAR function of the first order is

$$F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(y_{t-d}; c)]}, \gamma > 0, \quad (2)$$

while the ESTAR function is defined as

$$F(y_{t-d}; \gamma, c) = 1 - \exp[-\gamma(y_{t-d}; -c)^2], \gamma > 0. \quad (3)$$

The Absolute Error and Quadratic LSTAR Models

The transition variable can assume a lagged endogenous transition variable y_{t-d} as in Teräsvirta (1994), an exogenous variable, z_t in Adebile (2007), a linear and nonlinear function of lagged endogenous variables, $h(\tilde{y}_t; \alpha)$, which depends on a $(q \times 1)$ parameters vector or a linear time trend t which gives rise to a model with smoothly changing parameters (van Dijk et al., 2002). The absolute value of the random error term, t , from the initial AR(p) and the quadratic function at lag, d (that is $y_{t-d}^2 + b_{yt-d}$), are assumed to cause the transition from one regime to another in this paper. These are the AELSTAR and QLSTAR transition functions which give rise to AELSTAR and QLSTAR models, respectively, once they are substituted in the general STAR model.

The transition functions are

$$F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(|\epsilon_{t-d}| - c)]} - \frac{1}{2}, \gamma > 0 \quad (4)$$

and

$$F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp\left[-\gamma(y_{t-d}^2 + by_{t-d} - c)\right]} - \frac{1}{2}, \gamma > 0 \quad (5)$$

for AELSTAR and QLSTAR models, respectively. The parameters in the transition function are as defined earlier in (2) above. The $|\epsilon_{t-d}|$ is the absolute lagged error, and b is used to stabilize the transition variable y_{t-d} with $b = \pm 1$. The two transition functions above can be compared with the original LSTAR function of Teräsvirta (1994). Teräsvirta (1994) suggests subtracting 1/2 in order to ease the derivation of the test statistics (van Dijk et al., 2002), and this has been dropped in the proposed functions, Monte Carlo's procedures in this work, and the nonlinear parameters estimation.

In the AELSTAR function in (4), as the slope γ increases, $F(\epsilon_{t-d}; \gamma, c)$ quickly changes from 0 to 1. The switching between these regimes also depends on the contribution of the absolute lagged endogenous variable, $|\epsilon_{t-d}|$. Increasing the slope γ further ($\gamma \rightarrow \infty$) makes the transition function to the Self-Exciting Threshold Autoregressive (SETAR) in the context of STAR modelling, which is another variant of regime switching model. At $\gamma = 0$, $F(\epsilon_{t-d}; \gamma, c) = 0.5$, and hence the LSTAR(p) model becomes the linear Autoregressive AR(p) model. The QLSTAR function in (5) has similar properties to the LSTAR function by changing smoothly from 0 to 1 as γ increases and this also depends on the value of the quadratic expression which forces the LSTAR function to symmetric function. At $y_{t-d}^2 + by_{t-d} = c$, $F(c_{t-d}; \gamma, c) = 0.5$, hence the model is linear. Since $y_{t-d}^2 + by_{t-d} \geq c$, the quadratic inequality then has real roots, which implies the values of the transition variables y_{t-d} to be $y_{t-d} = -b \pm \sqrt{b^2 + 4c}$ for real roots using general quadratic function. For complex solutions (i.e. $b^2 < 4c$) there is a discontinuation in the transition from y_{t-d} to y_{t-d-1} . In the QLSTAR function in (5), the quadratic function in the transition variable as well as the slope parameter causes the transition to switch from one regime to the other, thereby absorbing the shock caused when the slope parameter, γ , suddenly increases or decreases. Unlike the traditional LSTAR function, the QLSTAR function produces a smooth symmetric realization.

The Shapes of the Proposed Transition Functions

Here, the original LSTAR and ESTAR functions are shown to be S and U shaped, respectively. Also, the Taylor series approximations (Teräsvirta, 1994; Escribano

& Jordá, 2001) to the functions are plotted alongside with the original LSTAR function.

The proposed transition functions, AELSTAR and QLSTAR, are considered here, and empirical analyses are given to investigate the shape and behavior of the functions. The value of the intercept, c , is fixed at 0.05, and that of the slope, γ , is varied as $\gamma = \{1, 10, 100\}$ for the original and proposed transition functions. Hypothetical values are then assigned for $|\epsilon_{t-d}|$ and y_{t-d} , and the corresponding values of $F(\cdot)$ are generated in both cases. The $F(\cdot)$ is then plotted on $|\epsilon_{t-d}|$ for the case of the AELSTAR function and on y_{t-d} for the case of the QLSTAR function. The proposed functions are then compared with the original LSTAR function.

Figure 1 is the graph of the LSTAR function plotted on values of $s_t = y_{t-d}$ for $\gamma = 1, 10$, and 100. At $\gamma = 1$, a straight line approximation is obtained; at $\gamma = 10$, there is a slower transition from $F(s_t = y_{t-d}) = 0$ to $F(s_t = y_{t-d}) = 1$ and therefore the shape of the LSTAR function is S. The transition is faster at $\gamma = 100$, and the shape of the LSTAR function at this point is mirrored Z. Both S and mirrored Z are asymmetric shapes. This reconfirms the asymmetric property of the LSTAR function. The point $y_{t-d} = c$ is the switch-point between the regime, and the graphs of the logistic function with various smoothness parameters meet at $F(c; \gamma, c) = 0.5$.

Figure 1. Behavior of the LSTAR function for different values of the slope parameter

Figure 2. Behavior of the AELSTAR function for different values of the slope parameter

Figure 3. Behavior of the QLSTAR function for different values of the slope parameter

Figure 2 is the AELSTAR function. At $\gamma = 1$, a wider V shape is obtained, and this becomes thinner as γ increases. The thinnest V shape obtained is for $\gamma = 100$. The V shapes obtained are reflective of the symmetric nature of the AELSTAR transition. Figure 3 is the QLSTAR function plotted on values of $s_t = y_{t-d}$, and shows that the widest U shape is obtained for $\gamma = 1$; this shape becomes thinner as γ increases. The U shape obtained is symmetric; hence the QLSTAR function is symmetrical.

Linearity Testing Procedures

The proposed LSTAR transition functions in (4) and (5) are defined with $1/2$ subtracted in order to ease the derivation of the linearity tests.

The AELSTAR Transition Function

The third order Taylor series approximation of the AELSTAR transition function around the null hypothesis $\gamma = 0$ is

$$\begin{aligned} F(\epsilon_{t-d}; \gamma, c) \approx & F(\epsilon_{t-d}; \gamma, c) \Big|_{\gamma=0} + \gamma \frac{\partial F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma \cdot 1!} \Big|_{\gamma=0} + \gamma^2 \frac{\partial^2 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^2 \cdot 2!} \Big|_{\gamma=0} \\ & + \gamma^3 \frac{\partial^3 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^3 \cdot 3!} \Big|_{\gamma=0} \end{aligned} \quad (6)$$

where

$$\begin{aligned} \text{a) } F(\epsilon_{t-d}; \gamma, c) &= \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-1} - \frac{1}{2}, F(\epsilon_{t-d}; \gamma, c) \Big|_{\gamma=0} = 0 \\ \text{b) } \frac{\partial F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma} &= \left((|\epsilon_{t-d}| - c) \exp[-\gamma(|\epsilon_{t-d}| - c)] \right) \left(\left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^2 \right)^{-1}, \\ \frac{\partial F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma} \Big|_{\gamma=0} &= \frac{1}{4} (|\epsilon_{t-d}| - c) \end{aligned}$$

$$\begin{aligned}
 \text{c) } \frac{\partial^2 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^2} &= 2(|\epsilon_{t-d}| - c)^2 \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-3} \left(\exp[-\gamma(|\epsilon_{t-d}| - c)] \right)^2 \\
 &\quad - (|\epsilon_{t-d}| - c)^2 \exp[-\gamma(|\epsilon_{t-d}| - c)] \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-2}, \\
 \frac{\partial^2 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^2 \cdot 2!} \Big|_{\gamma=0} &= 0 \\
 \frac{\partial^3 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^3} &= 6(|\epsilon_{t-d}| - c)^3 \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-4} \left(\exp[-\gamma(|\epsilon_{t-d}| - c)] \right)^3 \\
 &\quad - 2(|\epsilon_{t-d}| - c)^3 \exp[-\gamma(|\epsilon_{t-d}| - c)] \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-3} \\
 &\quad + (|\epsilon_{t-d}| - c)^3 \exp[-\gamma(|\epsilon_{t-d}| - c)] \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-2} \\
 &\quad - 2(|\epsilon_{t-d}| - c)^3 \left(\exp[-\gamma(|\epsilon_{t-d}| - c)] \right)^2 \left\{ 1 + \exp[-\gamma(|\epsilon_{t-d}| - c)] \right\}^{-3} \\
 \text{d) } \frac{\partial^3 F(\epsilon_{t-d}; \gamma, c)}{\partial \gamma^3 \cdot 3!} \Big|_{\gamma=0} &= \frac{1}{8} (|\epsilon_{t-d}| - c)^3
 \end{aligned}$$

Putting (a)-(d) above in (6),

$$\begin{aligned}
 F(\epsilon_{t-d}; \gamma, c) &= 0 + \gamma \frac{1}{4} \frac{|\epsilon_{t-d}| - c}{1!} + \frac{0}{2!} + \gamma^3 \frac{1}{8} \frac{1}{4} \frac{(|\epsilon_{t-d}| - c)^3}{3!} \\
 &= \frac{1}{4} \frac{\gamma (|\epsilon_{t-d}| - c)}{1!} + \frac{1}{48} \gamma^3 (|\epsilon_{t-d}| - c)^3 \\
 &= -\left(\frac{1}{4} c \gamma + \frac{1}{48} \gamma^3 c^3 \right) + \left(\frac{1}{4} \gamma + \frac{\gamma^3 c^3}{16} \right) |\epsilon_{t-d}| - \left(\frac{1}{16} c \gamma^3 |\epsilon_{t-d}|^2 \right) + \left(\frac{1}{48} \gamma^3 |\epsilon_{t-d}|^3 \right) \quad (7)
 \end{aligned}$$

The two-regime STAR model in (1) is equivalent to the LSTAR model

$$y_t = \Phi_1 \tilde{\mathbf{y}}_t^{(p)} + \Phi_2' \tilde{\mathbf{y}}_t^{(p)} F(\epsilon_{t-d}; \gamma, c) + \epsilon_t$$

with the transition function replaced by AELSTAR function. Substituting the approximated AELSTAR function in (7) in the AELSTAR model above gives

$$y_t = \phi_1' \tilde{y}_t^{(p)} + \phi_2' \tilde{y}_t^{(p)} \left[\begin{array}{c} -\left(\frac{1}{4}c\gamma + \frac{1}{48}\gamma^3 c^3\right) + \left(\frac{1}{4}\gamma + \frac{\gamma^3 c^3}{16}\right)|\epsilon_{t-d}| \\ + \left(\frac{1}{16}c\gamma^3 |\epsilon_{t-d}|^2\right) + \left(\frac{1}{48}\gamma^3 |\epsilon_{t-d}|^3\right) \end{array} \right] + \epsilon_t$$

which is equivalent to the auxiliary regression model

$$y_t = \gamma_0' \tilde{y}_t^{(p)} + \gamma_1' \tilde{y}_t^{(p)} |\epsilon_{t-d}| + \gamma_2' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^2 + \gamma_3' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^3 + \mathcal{G}_t . \quad (8)$$

Because γ_i ($i = 0, \dots, 3$) do not exist in real sense since they are parameters of the model, compute the corresponding estimates for them based on the regression model

$$y_t = \beta_0' \tilde{y}_t^{(p)} + \beta_1' \tilde{y}_t^{(p)} |\epsilon_{t-d}| + \beta_2' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^2 + \beta_3' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^3 + \mathcal{G}_t , \quad (9)$$

which is equivalent to

$$\mathcal{G}_t = \beta_0' \tilde{y}_t^{(p)} + \beta_1' \tilde{y}_t^{(p)} |\epsilon_{t-d}| + \beta_2' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^2 + \beta_3' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^3 + \varsigma_t , \quad (10)$$

where $\beta_i' \tilde{y}_t^{(p)}$, $i = 0, 1, 2, 3$ are functions of c and γ . The residuals $\mathcal{G}_t$ and ς_t are from the two auxiliary regressions, and are NID(0, σ^2). So the regression above is performed by regressing $\mathcal{G}_t$ on the series of regressors $\tilde{y}_t^{(p)}$ and $\tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^2$, ($i = 1, 2, 3$). At this stage, the existence of AELSTAR process depends on the parameters $\beta_1', \beta_2', \beta_3'$, which should not all be zeros. This further establishes the nonlinearity of the AELSTAR(p) model. So the nonlinear AELSTAR effect is tested with the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \quad (11)$$

tested against the standard Lagrange Multiplier (LM) test following the Teräsvirta decision rule. The test statistic has an asymptotic χ^2 distribution with $3(p+1)$ degrees of freedom and its F version defined as

$$F_3 = \frac{(\text{SSR}_0 - \text{SSR}_1)/3(p+1)}{\text{SSR}_1/(N-4(p+1))}, \quad (12)$$

where p is the order of the autoregression, N is the sample size, and SSR_0 and SSR_1 are the residual sum of squares from the linear and nonlinear specifications, respectively.

Testing nonlinearity in the proposed AELSTAR model is similar to that of LSTAR since the orders of y_{t-d} in the auxiliary model are up to 3. In a similar manner to the Teräsvirtä (1994) Procedure (TP), Escribano and Jordá (2001) Procedure (EJP) can also be generalized from the test for LSTAR nonlinearity which applied second-order Taylor series expansion of the transition function. In that case, obtained in a similar fashion, the auxiliary regression model is

$$y_t = \phi_1 \tilde{y}_t^{(p)} + \beta_1' \tilde{y}_t^{(p)} |\epsilon_{t-d}| + \beta_2' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^2 + \beta_3' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^3 + \beta_4' \tilde{y}_t^{(p)} (|\epsilon_{t-d}|)^4 + \tilde{\epsilon}_t \quad (13)$$

The null hypothesis of linearity is then given as

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0, \quad (14)$$

which is tested against the standard Lagrange Multiplier (LM) test using the EJP decision rule. The test statistic has an asymptotic χ^2 distribution with $4(p+1)$ degrees of freedom and its F version defined as

$$F_4 = \frac{(\text{SSR}_0 - \text{SSR}_1)/4(p+1)}{\text{SSR}_1/(N-5(p+1))}, \quad (15)$$

where p is the order of the autoregression, N is the sample size, and SSR_0 and SSR_1 are the residual sum of squares.

The QLSTAR Transition Function

For the QLSTAR transition function, the third order Taylor series approximation around the null hypothesis, $\gamma = 0$, is

$$\begin{aligned}
 F(\epsilon_{t-d}; \gamma, c) &\approx F(y_{t-d}; \gamma, c) \Big|_{\gamma=0} + \gamma \frac{\partial F(y_{t-d}; \gamma, c)}{\partial \gamma \cdot 1!} \Big|_{\gamma=0} \\
 &\quad + \gamma^2 \frac{\partial^2 F(y_{t-d}; \gamma, c)}{\partial \gamma^2 \cdot 1!} \Big|_{\gamma=0} + \gamma^3 \frac{\partial^3 F(y_{t-d}; \gamma, c)}{\partial \gamma^3 \cdot 1!} \Big|_{\gamma=0}
 \end{aligned} \tag{16}$$

where

$$\begin{aligned}
 \text{a) } F(y_{t-d}; \gamma, c) &= \left\{ 1 + \exp[-\gamma(y_{t-d} + by_{t-d} - c)] \right\}^{-1} - \frac{1}{2}, F(y_{t-d}; \gamma, c) \Big|_{\gamma=0} = 0 \\
 \text{b) } \frac{\partial F(y_{t-d}; \gamma, c)}{\partial \gamma} &= \left((y_{t-d}^2 + by_{t-d} - c) \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right) \\
 &\quad \times \left(\left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^2 \right)^{-1}, \\
 &\quad \frac{\partial F(y_{t-d}; \gamma, c)}{\partial \gamma} \Big|_{\gamma=0} = \frac{1}{4} (y_{t-d}^2 + by_{t-d} - c) \\
 \text{c) } \frac{\partial^2 F(y_{t-d}; \gamma, c)}{\partial \gamma^2} &= 2(y_{t-d}^2 + by_{t-d} - c)^2 \left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^{-3} \times \\
 &\quad \left(\exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] - (y_{t-d}^2 + by_{t-d} - c) \right)^2 \times \\
 &\quad \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^{-2}, \\
 &\quad \frac{\partial^2 F(y_{t-d}; \gamma, c)}{\partial \gamma^2} \Big|_{\gamma=0} = 0 \\
 \frac{\partial^3 F(y_{t-d}; \gamma, c)}{\partial \gamma^3} &= 6(y_{t-d}^2 + by_{t-d} - c)^3 \left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^{-4} \times \\
 &\quad \left(\exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right)^3 - 2(y_{t-d}^2 + by_{t-d} - c)^3 \times \\
 &\quad \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^{-3} \\
 &\quad + (y_{t-d}^2 + by_{t-d} - c)^3 \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \times \\
 \text{d) } &\quad \left\{ 1 + \exp[-\gamma(y_{t-d}^2 + by_{t-d} - c)] \right\}^{-2} - 2(y_{t-d}^2 + by_{t-d} - c)^3 \times
 \end{aligned}$$

$$\left. \frac{\partial^3 F(y_{t-d}; \gamma, c)}{\partial \gamma^3 \cdot 3!} \right|_{\gamma=0} = \frac{1}{8} (y_{t-d}^2 + by_{t-d} - c)^3$$

Putting steps (a-d) in (16),

$$\begin{aligned} F(y_{t-d}; \gamma, c) &= 0 + \gamma \frac{1}{4} \frac{y_{t-d}^2 + by_{t-d} - c}{1!} + \frac{0}{2} + \gamma^3 \frac{1}{8} \frac{\frac{1}{4} (y_{t-d}^2 + by_{t-d} - c)^3}{3!} \\ &= \frac{1}{4} \gamma (y_{t-d}^2 + by_{t-d} - c) + \frac{1}{48} \gamma^3 (y_{t-d}^2 + by_{t-d} - c)^3 \\ &= - \left(\frac{1}{4} c \gamma + \frac{1}{48} \frac{\gamma^3 c^3}{16} \right) + \left(\frac{1}{4} \gamma + \frac{\gamma^3 c^3}{16} \right) (y_{t-d}^2 + by_{t-d}) \\ &\quad - \left(\frac{1}{16} c \gamma^3 (y_{t-d}^2 + by_{t-d})^2 \right) + \left(\frac{1}{48} \gamma^3 (y_{t-d}^2 + by_{t-d})^3 \right) \end{aligned} \quad (17)$$

$$\begin{aligned} F(y_{t-d}; \gamma, c) &= -\frac{1}{48} c \gamma (c^2 \gamma^2 + 12) + \frac{1}{16} b \gamma (4 + c^2 \gamma^2) y_{t-d} \\ &\quad + \frac{1}{16} \gamma (4 + c^2 \gamma^2 - b^2 c \gamma^2) y_{t-d}^2 + \frac{1}{48} b \gamma^3 (b^2 - 6c) y_{t-d}^3 \\ &\quad + \frac{1}{16} b \gamma^3 (b^2 - c) y_{t-d}^4 + \frac{1}{16} b \gamma^3 y_{t-d}^5 + \frac{1}{48} \gamma^3 y_{t-d}^6 \end{aligned} \quad (18)$$

The two-regime STAR model in (1) is equivalent to the LSTAR model

$$y_t = \phi'_1 \tilde{y}_t^{(p)} + \phi'_2 \tilde{y}_t^{(p)} F(y_{t-d}; \gamma, c) + \epsilon_t$$

with the transition function replaced by QLSTAR function. Substituting the approximated QLSTAR function in (18) in the QLSTAR model above gives

LOGISTIC SMOOTH TRANSITION

$$y_t = \phi'_1 \tilde{y}_t^{(p)} + \phi'_2 \tilde{y}_t^{(p)} \left[-\frac{1}{48} c \gamma (c^2 \gamma^2 + 12) + \frac{1}{16} b \gamma (4 + c^2 \gamma^2) y_{t-d} \right. \\ \left. + \frac{1}{16} \gamma (4 + c^2 \gamma^2 - b^2 c \gamma^2) y_{t-d}^2 + \frac{1}{48} b \gamma^3 (b^2 - 6c) y_{t-d}^3 \right. \\ \left. + \frac{1}{16} \frac{1}{4} b \gamma^3 (b^2 - c) y_{t-d}^4 + \frac{1}{16} b \gamma^3 y_{t-d}^5 + \frac{1}{48} \gamma^3 y_{t-d}^6 \right] + \epsilon_t$$

which is equivalent to the auxiliary regression model

$$y_t = \gamma'_0 \tilde{y}_t^{(p)} + \gamma'_1 \tilde{y}_t^{(p)} y_{t-d} + \gamma'_2 \tilde{y}_t^{(p)} y_{t-d}^2 + \gamma'_3 \tilde{y}_t^{(p)} y_{t-d}^3 + \gamma'_4 \tilde{y}_t^{(p)} y_{t-d}^4 + \gamma'_5 \tilde{y}_t^{(p)} y_{t-d}^5 \\ + \gamma'_6 \tilde{y}_t^{(p)} y_{t-d}^6 + \mathcal{G}_t \quad (19)$$

Because γ_i ($i = 0, \dots, 6$) do not exist in real sense since they are parameters of the model, we therefore compute the corresponding estimates for them based on the regression model

$$y_t = \beta'_0 \tilde{y}_t^{(p)} + \beta'_1 \tilde{y}_t^{(p)} y_{t-d} + \beta'_2 \tilde{y}_t^{(p)} y_{t-d}^2 + \beta'_3 \tilde{y}_t^{(p)} y_{t-d}^3 + \beta'_4 \tilde{y}_t^{(p)} y_{t-d}^4 + \beta'_5 \tilde{y}_t^{(p)} y_{t-d}^5 \\ + \beta'_6 \tilde{y}_t^{(p)} y_{t-d}^6 + \mathcal{G}_t \quad (20)$$

which can be re-written as

$$\mathcal{G}_t = \beta'_0 \tilde{y}_t^{(p)} + \beta'_1 \tilde{y}_t^{(p)} y_{t-d} + \beta'_2 \tilde{y}_t^{(p)} y_{t-d}^2 + \beta'_3 \tilde{y}_t^{(p)} y_{t-d}^3 + \beta'_4 \tilde{y}_t^{(p)} y_{t-d}^4 \\ + \beta'_5 \tilde{y}_t^{(p)} y_{t-d}^5 + \zeta_t \quad (21)$$

where $\beta'_i y_t^{(p)}$, $i = 1, 2, 3$ are functions of b , c , and γ . Hence all the 6 terms are duly represented in the model. The model is QLSTAR if at least one of β'_i is not zero when the auxiliary regression is performed. So the regression above is performed by regressing $\mathcal{G}_t$ on the series of regressors $\tilde{y}_t^{(p)}$ and y_{t-d}^i , ($i = 1, \dots, 6$). Nonlinear QLSTAR is then tested based on the null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0 \quad (22)$$

which can be tested against standard Lagrange Multiplier (LM) test. The test statistic has an asymptotic χ^2 distribution with $6(p+1)$ degrees of freedom and its F version defined as

$$F_6 = \frac{(\text{SSR}_0 - \text{SSR}_1)/6(p+1)}{\text{SSR}_1/(N-7(p+1))} \quad (23)$$

The QLSTAR nonlinearity is then detected in the series once one of β_j , ($j = 1, \dots, 6$) is significantly different from zero in the auxiliary regression model in (21). The QLSTAR model is then specified when the level of significance of parameters β_1 , β_3 , and β_6 are higher than that of β_2 , β_4 , and β_6 . The counterpart, Quadratic ESTAR (QESTAR), is then chosen once there EW contrary results. Testing QLSTAR nonlinearity using Escribano and Jordá (2001) is much more straightforward. Linearity is established once the betas are zeroed. Once there is default in one, the decision rule of Escribano and Jordá (2001) is applied. Then apply the revised decision rule

$$\begin{aligned} H_{0\text{QLSTAR}} : \beta_1 = \beta_3 = \beta_5 = 0 \\ H_{0\text{QESTAR}} : \beta_2 = \beta_4 = \beta_6 = 0 \end{aligned} \quad (24)$$

Therefore, rejecting $H_{0\text{QLSTAR}}$ and failing to reject $H_{0\text{QESTAR}}$ points to QLSTAR model. Also, rejecting $H_{0\text{QESTAR}}$ and failing to reject $H_{0\text{QLSTAR}}$ suggests a QESTAR model. In addition, if the minimum p -value corresponds to F_6 , select QLSTAR and, if the minimum p -value corresponds to F_6 , select QESTAR model.

Monte Carlo (MC) Simulations

Due to the unavailability of the structural and distributional properties of the parameters of the STAR models as a result of unknown regularities conditions, Monte Carlo's simulation approach is then used to study the behaviors of some parameters in the model via the model selection procedures. The accuracy of the TP and EJP approaches in selecting between the two types of LSTAR models is then examined. The correct selection rate is reported as a proportion per 1000 replications for which linearity was first rejected at 95% confidence level. The robustness of the selection procedures is also tested by considering non-zero threshold values, $c \neq 0$ that make the model to be asymmetric.

LOGISTIC SMOOTH TRANSITION

The program for the simulation is set under the OxGAUSS engine. Most nonlinear simulation and estimations are performed under GAUSS machines, particularly with Ox, since this helps to perform the simulations at a faster rate when compared to other software or ordinary GAUSS software. (Lin, 2001; Chan and McAleer, 2001; 2003). We first consider using the STAR Data Generating Process (DGP) used in Granger and Teräsvirta (1993), Teräsvirta, Lin, and Granger (1993), Teräsvirta (1994), Escribano, Franses, and van Dijk (1998), Escribano and Jordá (2001), Lopes and Salazar (2006), and Adebile (2007):

$$y_t = 1.800y_{t-1} - 1.060y_{t-2} (\phi_{20} - 0.900y_{t-1} + 0.795y_{t-2}) F(s_t; \gamma, c) + \epsilon_t \quad (25)$$

where $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for the STAR specification. The values for ϕ_{20} are set as $\phi_{20} = \{0.0, 0.2\}$. The STAR DGP in (25) is suggested due to the fact that it has been used over and over by researchers. Probing into the DGP, when the transition function $F(s_t; \gamma, c) = 0$, the system is in the lower regime and gives a linear AR model with the characteristic equation $1.06B^2 - 1.80B + 1 = 0$, which has roots $|0.849 \pm 0.472i|$ in the complex plane and the modulus of the complex root is 0.971. The real root is 0.849, which is less than unity; this implies nonstationarity of the realized time series process. This may cause an explosion unless stationarity is imposed on the nonlinear part of the DGP, and y_t will have the tendency to adjust to more stable parts of the state space.

Teräsvirta (1994) and Escribano et al. (1998) supported this assertion. Also, when $F(s_t; \gamma, c) = 1$, the system is in the upper regime and gives the AR model $y_t = \phi_{20} + 0.900y_{t-1} - 0.265y_{t-2} + \epsilon_t$. Because ϕ_{20} has values 0.0 and 0.2, we therefore have two characteristic equations: $0.265B^2 - 0.900B + 1 = 0$ and $0.265B^2 - 0.900B + 0.8 = 0$, which give the complex roots $|1.698 \pm 0.943i|$ and $|1.698 \pm 0.368i|$ with the moduli of 1.942 and 1.737, respectively. These roots lie outside the unit circle, and this will control the system to realize stationary series. These are the motives for using the GDP. It is obvious to see that the properties of time series to be generated will then depend on the relative magnitude of ϕ_{20} and c as they jointly determine the value of y_t and its instability and stability (Escribano et al., 1998).

Occasionally, the software simulates values $y_t < c$ which makes the linear AR model to be explosive, and this point is not common in real life situation (Escribano & Jordá, 2001). A burn-in of 100 observations shields the experiments against a potential dependence on starting values and, after making allowance to discard these, we proceed in the simulations. It is also noted that higher values of the slope parameter, γ , will cause the power of the two selection procedures to be closed to

each other, and this makes the nonlinearity to be sharp and significant. Therefore this will make the discrimination between the two models to be close to each other (Escribano & Jordá, 2001). So we also consider setting $\gamma = \{1, 10, 100\}$ and $c = \{0.0, 0.2, 0.5, 1.0\}$ in the transition functions.

The DGPs are LSTAR with the transition functions

$$F(s_t; \gamma, c) \equiv F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(y_{t-d} - c)]} \quad (26)$$

$$F(\epsilon_t; \gamma, c) \equiv F(\epsilon_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(\epsilon_{t-d} - c)]} \quad (27)$$

$$F(s_t; \gamma, c) \equiv F(y_{t-d}; \gamma, c) = \frac{1}{1 + \exp[-\gamma(y_{t-d}^2 + \epsilon_{t-d} - c)]} \quad (28)$$

Relative frequencies of the models specification are computed and presented on every 1000 replications and across sample sizes $N = \{50, 100, 200, 500, 1000\}$. The cells report the number of times a variant of LSTAR model is selected at 5% nominal significant level in every 1000 replications.

The results are presented in Tables 1-6 below. From Table 1, ϕ_{20} , the nonlinear component of the DGP realizes real values though the linear component is still complex. The QLSTAR model compares favorably well with the LSTAR based on the specification tests, TP and EJP. Improvement in selection frequencies is seen as sample sizes increase, therefore specification of models is consistent with sample size. TP outperforms EJP when the threshold, c , is marginally different from zero (say $c = 0.2$). Comparison of the models shows that the proposed AELSTAR and QLSTAR models seem to realize lower frequencies when compared with the original LSTAR model. This implies that the data generated by the proposed models resembled ESTAR, which is a symmetric series, and this may be clearer when the variation in the series is increased.

With increased standard deviation $\sigma = 0.1$, more selection frequencies of selection are computed for the variants of the model. Unlike the results in Table 1, here the frequencies are computed for the heteroscedastic versions of the models. This implies that the level of variations in the time series a direct the specification of the models. In the selection of variants of LSTAR models, the TP dominates EJP in discriminating between the two types of STAR models as indicated in the frequencies computed based on the two procedures for each of the LSTAR models. Also, the frequencies computed by the variants of LSTAR are smaller than the values realized by the original LSTAR model, which is indication that the

AELSTAR and QLSTAR models detect symmetric time series. The selection frequencies imply higher frequencies computed for the ESTAR models.

In Table 2, the TP outperforms the EJP at different simulation points. TP is also seen to be more consistent with increasing sample sizes. Matrix inversion problems were also reported during the simulation due to lower variation in the series. Increasing the parameter $\phi_{20} = 0.0$ to $\phi_{20} = 0.2$ has caused an increase in the model selection frequencies. Since this parameter directs the stationarity level of the realized series, it then implies that stationarity has a relationship with nonlinearity of the series. For model comparison, the results here follow that of Table 1, only that the improvement in the symmetry as indicated by the proposed transition functions is not as significant as the results given in Table 1.

From Table 3, with increased in transition speed $\gamma = 10$, nonlinearity is clearer due to more points that are computed for frequencies of selection. QLSTAR model competes favorably well with the LSTAR. Increase in transition speed also led to the computation of the GARCH component in the model. This implies that the GARCH effect is much felt in the series when nonlinearity is sharper. EJP is more consistent in selecting the variants of the LSTAR model up to $c = 0.5$ and $N = 1000$. TP is consistent up to $c = 0.2$ and $N = 1000$ for LSTAR and QLSTAR models. EJP is seen to dominate TP in LSTAR and QLSTAR versions of the models, whereas in AELSTAR models, TP dominates EJP.

In Table 4, both TP and EJP compete well in selecting variants of AELSTAR and QLSTAR models. More frequencies are computed for the AELSTAR model, and this favors the selection of the LSTAR model. The proposed models realized symmetric time series as indicated in the lower frequencies realized for the LSTAR models by the two specification procedures.

With nonlinearity further increased to transition speed $\gamma = 100$, more selection frequencies are generated. EJP is more sensitive to selection of LSTAR and QLSTAR models, while TP is sensitive to selecting only the AELSTAR model. In selecting the LSTAR model, only EJP is consistent with sample sizes throughout. The results indicated that symmetric time series are realized by the proposed models as indicated by the lower frequencies computed for the LSTAR models which is known to be asymmetric.

Table 1. Selection frequencies of variants of LSTAR model at $\gamma = 1$, $\phi_{20} = 0.0$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.0	500	-	-	0.740	0.480	-	-	0.698	0.420	-	-	0.750	0.500
0.0	1000	-	-	0.750	0.650	-	-	0.641	0.590	-	-	0.690	0.603
0.0	200	0.538	0.410	0.762	0.655	0.474	0.421	0.579	0.395	0.538	0.410	0.659	0.622
0.0	500	0.707	0.537	0.896	0.786	0.683	0.537	0.615	0.538	0.707	0.537	0.825	0.748
0.0	10000	0.744	0.651	0.952	0.915	0.722	0.583	0.722	0.500	0.732	0.659	0.880	0.877
0.2	50	-	-	0.700	0.560	-	-	0.676	0.486	-	-	0.731	0.519
0.2	100	-	-	0.754	0.692	-	-	0.641	0.590	-	-	0.758	0.677
0.2	200	0.486	0.541	0.787	0.719	0.421	0.447	0.394	0.424	0.486	0.568	0.648	0.625
0.2	500	0.721	0.581	0.912	0.814	0.691	0.524	0.643	0.524	0.711	0.556	0.837	0.770
0.2	1000	0.755	0.667	0.963	0.940	0.735	0.588	0.737	0.527	0.733	0.644	0.886	0.886
0.5	50	-	-	0.712	0.577	-	-	0.667	0.615	-	-	0.714	0.571
0.5	100	-	-	0.857	0.814	-	-	0.778	0.667	-	-	0.754	0.783
0.5	200	0.525	0.550	0.803	0.752	0.528	0.500	0.455	0.424	0.537	0.537	0.679	0.696
0.5	500	0.694	0.612	0.920	0.877	0.632	0.500	0.565	0.543	0.674	0.587	0.819	0.808
0.5	1000	0.740	0.580	0.972	0.967	0.733	0.556	0.732	0.561	0.698	0.566	0.888	0.907
1.0	50	-	-	0.697	0.643	-	-	0.600	0.600	-	-	0.717	0.583
1.0	100	-	-	0.828	0.862	-	-	0.771	0.686	-	-	0.691	0.742
1.0	200	0.561	0.463	0.824	0.818	0.548	0.387	0.455	0.394	0.538	0.436	0.684	0.759
1.0	500	0.644	0.511	0.952	0.945	0.649	0.486	0.579	0.474	0.652	0.565	0.781	0.842
1.0	1000	0.685	0.589	0.983	0.986	0.692	0.500	0.679	0.472	0.685	0.589	0.851	0.915

LOGISTIC SMOOTH TRANSITION

Table 2. Selection frequencies of variants of LSTAR model at $\gamma = 1$, $\phi_{20} = 0.2$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.0	500	-	-	0.721	0.465	-	-	0.658	0.421	-	-	0.589	0.464
0.0	1000	-	-	0.768	0.518	-	-	0.686	0.486	-	-	0.726	0.493
0.0	200	0.595	0.500	0.778	0.583	0.526	0.526	0.568	0.522	0.707	0.463	0.745	0.480
0.0	500	0.692	0.385	0.893	0.508	0.700	0.550	0.707	0.683	0.721	0.372	0.909	0.514
0.0	10000	0.750	0.591	0.933	0.601	0.767	0.488	0.604	0.583	0.700	0.600	0.973	0.496
0.2	50	-	-	0.710	0.421	-	-	0.667	0.500	-	-	0.611	0.429
0.2	100	-	-	0.737	0.491	-	-	0.737	0.500	-	-	0.756	0.488
0.2	200	0.625	0.550	0.805	0.61	0.421	0.579	0.459	0.595	0.659	0.561	0.744	0.529
0.2	500	0.732	0.390	0.928	0.582	0.683	0.561	0.659	0.705	0.690	0.357	0.907	0.534
0.2	1000	0.745	0.588	0.944	0.628	0.800	0.450	0.761	0.478	0.708	0.631	0.965	0.510
0.5	50	-	-	0.761	0.522	-	-	0.703	0.432	-	-	0.644	0.508
0.5	100	-	-	0.772	0.474	-	-	0.778	0.417	-	-	0.774	0.560
0.5	200	0.657	0.514	0.826	0.616	0.545	0.545	0.529	0.529	0.625	0.475	0.739	0.556
0.5	500	0.745	0.426	0.914	0.686	0.628	0.535	0.698	0.721	0.698	0.528	0.900	0.537
0.5	1000	0.732	0.500	0.947	0.736	0.750	0.523	0.708	0.438	0.774	0.613	0.955	0.557
1.0	50	-	-	0.780	0.560	-	-	0.567	0.600	-	-	0.787	0.591
1.0	100	-	-	0.736	0.605	-	-	0.765	0.618	-	-	0.739	0.635
1.0	200	0.634	0.537	0.819	0.624	0.500	0.600	0.441	0.559	0.636	0.477	0.727	0.556
1.0	500	0.659	0.500	0.935	0.791	0.684	0.605	0.610	0.488	0.642	0.434	0.871	0.595
1.0	1000	0.696	0.551	0.980	0.858	0.719	0.386	0.722	0.463	0.653	0.431	0.920	0.643

Table 3. Selection frequencies of variants of LSTAR model at $\gamma = 10$, $\phi_{20} = 0.0$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.0	500	-	-	0.922	0.962	-	-	0.698	0.4199	-	-	0.750	0.500
0.0	1000	-	-	0.973	0.994	-	-	9		-	-	0.690	0.603
0.0	200	0.918	0.863	0.995	1.000	0.575	0.450	0.641	0.590	0.881	0.852	0.659	0.622
0.0	500	0.986	0.973	1.000	1.000	0.684	0.553	0.579	0.395	0.983	0.970	0.825	0.748
0.0	10000	1.000	0.993	1.000	1.000	0.522	0.478	0.615	0.538	0.999	0.992	0.880	0.870
0.2	50	-	-	0.870	0.966	-	-	0.722	0.500	-	-	0.731	0.519
0.2	100	0.897	0.932	0.928	0.994	0.585	0.463	0.676	0.486	0.875	0.913	0.758	0.677
0.2	200	0.959	0.982	0.979	1.000	0.629	0.514	0.641	0.590	0.941	0.974	0.648	0.625
0.2	500	0.996	1.000	1.000	1.000	0.608	0.431	0.394	0.424	0.993	0.997	0.837	0.770
0.2	1000	1.000	1.000	1.000	1.000	0.612	0.531	0.643	0.524	1.000	1.000	0.886	0.886
0.5	50	-	0.964	0.773	0.887	0.700	0.550	0.737	0.526	-	0.944	0.714	0.571
0.5	100	-	0.997	0.865	0.961	0.680	0.480	0.667	0.615	-	0.983	0.754	0.783
0.5	200	-	1.000	0.936	0.992	0.700	0.550	0.778	0.667	-	0.998	0.679	0.696
0.5	500	-	1.000	0.997	1.000	0.683	0.542	0.455	0.524	-	1.000	0.819	0.808
0.5	1000	-	1.000	1.000	1.000	0.642	0.512	0.565	0.543	-	1.000	0.888	0.907
1.0	50	-	1.000	0.604	0.894	0.692	0.481	0.732	0.561	-	0.997	0.717	0.583
1.0	100	-	0.999	0.664	0.960	0.735	0.490	0.600	0.600	-	1.000	0.691	0.742
1.0	200	-	1.000	0.763	0.998	0.667	0.471	0.771	0.686	-	1.000	0.684	0.759
1.0	500	-	1.000	0.914	1.000	0.664	0.502	0.455	0.394	-	1.000	0.781	0.842
1.0	1000	-	-	1.000	1.000	0.673	0.525	0.579	0.474	-	1.000	0.851	0.915

LOGISTIC SMOOTH TRANSITION

Table 4. Selection frequencies of variants of LSTAR model at $\gamma = 10$, $\phi_{20} = 0.2$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.	500	-	-	0.776	0.531	-	-	0.604	0.566	-	-	0.723	0.409
0.	1000	-	-	0.843	0.585	-	-	0.627	0.627	-	-	0.758	0.532
0.	200	-	-	0.842	0.614	-	-	0.577	0.627	-	-	0.832	0.589
0.	500	-	-	0.854	0.776	0.830	0.616	0.731	0.737	-	-	0.821	0.758
0.	1000	-	-	0.888	0.904	0.830	0.616	0.809	0.789	-	-	0.843	0.870
0.	50	-	-	0.775	0.598	-	-	0.484	0.453	-	-	0.706	0.557
0.	100	-	-	0.912	0.616	0.605	0.579	0.600	0.537	-	-	0.904	0.565
0.	200	-	-	0.974	0.691	0.667	0.667	0.411	0.492	-	-	0.963	0.586
0.	500	-	-	0.999	0.820	0.580	0.540	0.475	0.660	-	-	0.989	0.709
0.	1000	0.754	0.492	1.000	0.923	0.661	0.597	0.480	0.724	-	-	0.993	0.787
0.	50	0.794	0.815	0.724	0.686	0.661	0.597	0.620	0.480	-	-	0.745	0.670
0.	100	0.911	0.931	0.810	0.758	0.667	0.453	0.703	0.520	-	-	0.849	0.739
0.	200	0.974	0.982	0.910	0.855	0.705	0.558	0.696	0.532	-	-	0.933	0.843
0.	500	0.999	0.999	0.988	0.964	0.679	0.544	0.672	0.523	0.951	0.583	0.969	0.961
0.	1000	1.000	1.000	0.988	0.964	0.648	0.516	0.493	0.658	0.991	0.659	0.991	0.990
1.	50	0.678	0.938	-	0.722	0.692	0.481	0.679	0.472	0.338	0.779	0.612	0.646
1.	100	0.682	0.982	-	0.780	0.740	0.500	0.761	0.522	0.213	0.851	0.647	0.718
1.	200	0.740	1.000	-	0.887	0.667	0.471	0.667	0.504	0.085	0.949	0.697	0.744
1.	500	0.806	1.000	-	0.970	0.664	0.502	0.674	0.522	0.009	0.996	0.766	0.854
1.	1000	-	-	-	0.996	0.672	0.525	0.663	0.521	0.000	1.000	0.830	0.919

Table 5. Selection frequencies of variants of LSTAR model at $\gamma = 100$, $\phi_{20} = 0.0$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.0	500	-	-	0.941	0.989	-	-	0.600	0.511	-	-	0.889	0.985
0.0	1000	-	-	0.975	1.000	-	-	0.500	0.455	-	-	0.912	1.000
0.0	200	0.994	1.000	0.999	1.000	-	-	0.698	0.581	0.994	1.000	0.921	1.000
0.0	500	1.000	1.000	1.000	1.000	-	-	0.611	0.389	1.000	1.000	0.976	1.000
0.0	1000	1.000	1.000	1.000	1.000	0.500	0.444	0.725	0.500	1.000	1.000	0.995	1.000
0.2	50	-	0.983	0.885	0.962	0.692	0.481	0.672	0.605	0.686	0.885	0.818	0.977
0.2	100	-	0.968	0.949	0.991	0.735	0.510	0.777	0.658	0.783	0.968	0.853	0.994
0.2	200	0.830	0.999	0.987	1.000	0.660	0.440	0.757	0.640	0.909	0.995	0.883	1.000
0.2	500	-	1.000	1.000	1.000	0.715	0.548	0.699	0.606	0.991	1.000	0.951	1.000
0.2	1000	-	1.000	1.000	1.000	0.635	0.501	0.650	0.554	1.000	1.000	0.984	1.000
0.5	50	-	0.994	0.780	0.868	0.692	0.481	0.692	0.481	0.498	0.980	0.723	0.907
0.5	100	-	1.000	0.898	0.960	0.735	0.510	0.735	0.510	0.576	1.000	0.750	0.962
0.5	200	-	1.000	0.967	0.994	0.660	0.440	0.660	0.441	0.739	1.000	0.759	0.997
0.5	500	-	1.000	0.999	1.000	0.768	0.643	0.776	0.592	0.900	1.000	0.827	1.000
0.5	1000	-	1.000	1.000	1.000	1.000	0.000	0.663	0.588	0.978	1.000	0.894	1.000
1.0	50	-	0.998	0.657	0.901	0.692	0.481	0.692	0.481	0.414	0.995	0.580	0.854
1.0	100	-	1.000	0.779	0.973	0.735	0.510	0.735	0.510	0.515	1.000	0.568	0.936
1.0	200	-	1.000	0.918	0.999	0.660	0.440	0.660	0.440	0.576	1.000	0.512	0.990
1.0	500	--	1.000	0.993	1.000	0.768	0.643	0.768	0.643	0.774	1.000	0.492	1.000
1.0	1000	-	1.000	1.000	1.000	1.000	0.000	0.000	1.000	0.884	1.000	0.488	1.000

LOGISTIC SMOOTH TRANSITION

Table 6. Selection frequencies of variants of LSTAR model at $\gamma = 100$, $\phi_{20} = 0.2$ for $\epsilon_t \sim N(0, 0.02^2)$ and $\epsilon_t \sim N(0, 0.10^2)$ for STAR Specifications

c	N	LSTAR				AELSTAR				QLSTAR			
		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$		$\epsilon_t \sim N(0, 0.02^2)$		$\epsilon_t \sim N(0, 0.10^2)$	
		TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP	TP	EJP
0.0	500	-	-	0.617	0.532	-	-	0.553	0.426	-	-	0.617	0.532
0.0	1000	-	-	0.569	0.510	-	-	0.447	0.523	-	-	0.549	0.490
0.0	200	-	-	0.733	0.550	0.917	0.421	0.679	0.491	-	-	0.705	0.557
0.0	500	-	-	0.763	0.700	0.984	0.515	0.746	0.380	-	-	0.750	0.688
0.0	1000	-	-	0.850	0.796	0.998	0.484	0.890	0.378	-	-	0.847	0.802
0.2	50	-	-	0.751	0.643	0.692	0.491	0.710	0.571	-	-	0.646	0.653
0.2	100	-	-	0.905	0.659	0.735	0.510	0.777	0.642	-	-	0.854	0.637
0.2	200	-	-	0.979	0.679	0.660	0.440	0.747	0.657	-	-	0.933	0.681
0.2	500	-	-	0.993	0.818	0.715	0.548	0.699	0.604	-	-	0.960	0.829
0.2	1000	-	-	1.000	0.908	0.635	0.501	0.638	0.550	-	-	0.966	0.919
0.5	50	0.744	0.747	0.731	0.666	0.692	0.481	0.692	0.481	-	-	0.743	0.644
0.5	100	0.835	0.856	0.837	0.751	0.735	0.510	0.735	0.510	-	-	0.838	0.728
0.5	200	0.923	0.962	0.924	0.835	0.660	0.440	0.660	0.440	-	-	0.912	0.824
0.5	500	0.987	0.998	0.994	0.960	0.768	0.643	0.776	0.596	-	-	0.944	0.943
0.5	1000	0.999	1.000	1.000	0.995	1.000	0.000	0.670	0.592	-	-	0.980	0.999
1.0	50	-	0.951	0.535	0.739	0.692	0.481	0.692	0.481	-	0.807	0.618	0.655
1.0	100	-	0.997	0.573	0.824	0.735	0.510	0.735	0.510	-	0.937	0.663	0.717
1.0	200	-	1.000	0.645	0.905	0.660	0.440	0.660	0.440	-	0.990	0.723	0.765
1.0	500	-	1.000	0.774	0.969	0.768	0.643	0.768	0.643	-	1.000	0.776	0.838
1.0	1000	-	1.000	0.865	0.996	1.000	0.000	0.000	1.000	-	1.000	0.822	0.899

Conclusion

Two variants of LSTAR models were proposed which realized time series data with improved symmetry over the existing LSTAR model, which is known to be asymmetric. These proposed models were the AELSTAR and QLSTAR models. Monte Carlo experiment was set up, and selections of the nonlinear model was based on Teräsvirta Procedure (TP) and Escribano and Jordá Procedure (EJP). The two selection procedures selected models at frequencies lower than that of the LSTAR model, and this implied more frequencies of selection of symmetric variants of the model. The selection frequency increased as nonlinearity power in the model and sample sizes increased. This work therefore presents the AELSTAR and QLSTAR models as better alternatives to the existing LSTAR model in empirical economic and financial modeling.

References

- Acemoglu, D., & Scott, A. (1994). Asymmetries in the cyclical behaviour of UK labour markets. *The Economic Journal*, 104(427), 1303-1323. doi: 10.2307/2235450
- Adebile, O. A. (2007). *Specification, estimation and evaluation of smooth transition autoregressive models* (Unpublished doctoral thesis). Department of Statistics, University of Ibadan, Ibadan, Nigeria.
- Aslanidis, N., Osborn, D. R., & Sensier, M. (2002). Smooth Transition Regression Models in UK Stock Returns (Working Paper). Retrieved from <http://fmwww.bc.edu/RePEc/res2002/Aslanidis.pdf>
- Baum, C. F., Caglayan, M., & Barkoulas, J. T. (1998). Nonlinear adjustment to purchasing power parity in the post-Bretton Woods era (Working Paper Number 404). Retrieved from <http://fmwww.bc.edu/ec-p/wp404.pdf>
- Chan, K., & McAleer, M. (2001). Maximum likelihood estimation of STAR and STARGARCH models: A Monte Carlo analysis. *Journal of Applied Econometrics*, 175(5), 509-534. doi: 10.1002/jae.686
- Chan, K., & McAleer, M. (2003). On the structure, asymptotic theory and application of STAR-GARCH models (Unpublished Manuscript).

- Chan, K., & Tong, H. (1986). On estimating thresholds in autoregressive models. *Journal of Time Series Analysis*, 7(3), 179-190. doi: 10.1111/j.1467-9892.1986.tb00501.x
- Dueker, M., Sola, M., & Spagnolo, F. (2006). Contemporaneous threshold autoregressive models: Estimation, testing and forecasting (Working Paper 2003-024C). Retrieved from <https://research.stlouisfed.org/wp/2003/2003-024.pdf>
- Escribano, A., Franses, P. H., & van Dijk, D. (1998). Nonlinearities and Outliers: Robust Specification of STAR Models (Econometric Institute Research Report 9832/A).
- Escribano, A., & Jordá, O. (1999). Improved testing and specification of smooth transition regression models. In P. Rothman (Ed.), *Nonlinear time series analysis of economic and financial data* (pp. 289-319). Boston, MA: Kluwer.
- Escribano, A., & Jordá, O. (2001). Testing nonlinearity: Decision rules for selecting between logistic and exponential STAR models. *Spanish Economic Review*, 3(3), 193-209. doi: 10.1007/PL00011442
- Granger, C. W. J., & Teräsvirta, T. (2006). *Modelling nonlinear economic relationships*. Shanghai, China: Shanghai University of Finance and Economics Press.
- Liew, V. K. S., Ahmad, Z. B., & Sie-Hoe, L. (2002). Forecasting performance of logistic smooth transition autoregressive exchange rate model: The original and reparameterised versions (MPRA Paper 511).
- Lin, K-P. (2001). *Computational econometrics: GAUSS programming for econometricians and financial analysts*. Los Angeles, CA: ETEXT Publishing.
- Lopes, H. F., & Salazar, E. (2006). Time series mean level and stochastic volatility modelling by smooth transition autoregressions: A Bayesian approach. *Advances in Econometrics*, 20(2), 225-238. doi: 10.1016/S0731-9053(05)20028-2
- Öcal, N. (2000). Nonlinear models for U.K. macroeconomic time series. *Studies in Nonlinear Dynamics & Econometrics*, 4(3), 123-135. doi: 10.2202/1558-3708.1061
- Siliverstovs, B. (2005). The Bi-parameter smooth transition autoregressive model. *Economics Bulletin*, 3(23), 1-11. Retrieved from <http://www.accessecon.com/pubs/eb/default.aspx?topic=Abstract&PaperID=EB-05C50003>
- Skalin, J., & Teräsvirta, T. (1996). Another look at Swedish business cycles (Working Paper Series in Economic and Finance, Number 130).

- Skalin, J., & Teräsvirta, T. (1998). Modelling asymmetries in unemployment rates (Working Paper Series in Economic and Finance, Number 262).
- Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, 89(425), 208-218. doi: [10.1080/01621459.1994.10476462](https://doi.org/10.1080/01621459.1994.10476462)
- Teräsvirta, T., & Anderson, H. M. (1992). Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics*, 7(Supplement S1), S119-S136. doi: [10.1002/jae.3950070509](https://doi.org/10.1002/jae.3950070509)
- Teräsvirta, T., Lin, C-F. J., & Granger, C. W. J. (1993). Power of the neural network linearity test. *Journal of Time Series Analysis*, 14(2), 209-220. doi: [10.1111/j.1467-9892.1993.tb00139.x](https://doi.org/10.1111/j.1467-9892.1993.tb00139.x)
- Teräsvirta, T., van Dijk, D., & Medeiros, M. C. (2005). Linear models, smooth transition autoregressions, and neural networks for forecasting of macroeconomic time series: A re-examination. *International Journal of Forecasting*, 21(4), 755-774. doi: [10.1016/j.ijforecast.2005.04.010](https://doi.org/10.1016/j.ijforecast.2005.04.010)
- Tsay, R. S. (2005). Analysis of financial time series (2nd ed.). Hoboken, NJ: Wiley-Interscience.
- van Dijk, D. (1999). Smooth transition models: Extensions and outlier robust inference (Doctoral thesis). Thela Thesis, Amsterdam. Retrieved from <http://hdl.handle.net/1765/1856>
- van Dijk, D., Teräsvirta, T., & Franses, P. H. (2002). Smooth transition autoregressive models—A survey of recent developments. *Econometric Reviews*, 21(1), 1-47. doi: [10.1081/ETC-120008723](https://doi.org/10.1081/ETC-120008723)
- Woodward, G., & Anderson, H. M. (2009). Does beta react to market conditions? Estimates of ‘bull’ and ‘bear’ betas using a nonlinear market model with an endogenous threshold parameter. *Quantitative Finance*, 9(8), 913-924. doi: [10.1080/14697680802595643](https://doi.org/10.1080/14697680802595643)

Liu-Type Logistic Estimators with Optimal Shrinkage Parameter

Yasin Asar

Necmettin Erbakan University
Konya, Turkey

Multicollinearity in logistic regression affects the variance of the maximum likelihood estimator negatively. In this study, Liu-type estimators are used to reduce the variance and overcome the multicollinearity by applying some existing ridge regression estimators to the case of logistic regression model. A Monte Carlo simulation is given to evaluate the performances of these estimators when the optimal shrinkage parameter is used in the Liu-type estimators, along with an application of real case data.

Keywords: Logistic regression, multicollinearity, maximum likelihood, MSE, Liu-type estimator

---

Introduction

It is a very common problem to deal with highly intercorrelated explanatory variables. Especially in economics and in other applied research areas, the variables used in the multiple linear regression models are collinear. This problem is called multicollinearity. There are some results of the multicollinearity problem such as having inflated variance scores and instable estimations of the parameters when the ordinary least square (OLS) estimator is used. Similarly, in the logistic regression model, if the maximum likelihood method is used, these drawbacks occur at all. Also, one cannot obtain decisive answers to the related questions in both of the models.

There are some methods to deal with this problem. One method is to use ridge regression, first introduced by Hoerl and Kennard (1970). The other methods are to use Liu or Liu-type estimators proposed by Liu (1993) and Liu (2003) respectively.

These methods have been applied in the case of multiple linear models. However, there is not much attention paid to the multicollinearity problem in the case of logistic regression. Månsson and Shukur (2011), Kibria, Månsson, and

---

Mr. Asar is an Assistant Professor in the Department of Mathematics-Computer Science. Email him at: yasar@konya.edu.tr or at yasinasar@hotmail.com.

Shukur (2012), and Inan and Erdogan (2013) are some exceptions. In the first two studies, the authors used some early defined ridge estimators in the logistic regression model. In the last one, the authors applied Liu-type estimators given in Liu (1993) to the logistic model as well.

There is another Liu-type estimator defined by Huang (2012) for the logistic regression model. The author explained the theoretical advantages of this estimator and gave some comparisons. In this study, we use this estimator with optimal shrinkage parameter and some existing ridge parameters in order to make a simulation study to see the performance of these estimators in the logistic regression model.

Methodology

Consider the binary logistic regression model, a widely used method in statistical analysis, such that the dependent variable is $\text{Be}(\mathbf{P})$ where $\mathbf{P} = \frac{e^{\mathbf{X}\boldsymbol{\beta}}}{1 + e^{\mathbf{X}\boldsymbol{\beta}}}$ such that $\mathbf{X}$ is the design matrix of order $n \times (p + 1)$, p is the number of explanatory variables, and $\boldsymbol{\beta}$ is the coefficient vector of order $(p + 1) \times 1$. The most commonly applied method of estimating $\boldsymbol{\beta}$ is the maximum likelihood estimation (MLE) method.

One can compute the MLE of $\boldsymbol{\beta}$ by using the iteratively re-weighted least square algorithm as follows:

$$\hat{\boldsymbol{\beta}}_{\text{MLE}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} \mathbf{X}'\mathbf{W}\hat{\mathbf{z}}, \quad (1)$$

where $\mathbf{W} = \text{diag}(P_i(1 - P_i))$ and $\hat{z}_i = \log(P_i) + \frac{y_i - P_i}{P_i(1 - P_i)}$ is the i^{th} element of the vector $\hat{\mathbf{z}}$, $i = 1, 2, \dots, n$.

The mean square error (MSE) of the MLE is given as follows:

$$\text{MSE}(\hat{\boldsymbol{\beta}}_{\text{MLE}}) = E \left[(\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta})' (\hat{\boldsymbol{\beta}}_{\text{MLE}} - \boldsymbol{\beta}) \right] = \text{tr}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1} = \sum_{j=1}^{p+1} \frac{1}{\lambda_j}, \quad (2)$$

where λ_j is the j^{th} eigenvalue of the matrix $\mathbf{X}'\mathbf{W}\mathbf{X}$.

If some of the eigenvalues are small (close to zero), then the asymptotic variance of MLE becomes inflated. In other words, multicollinearity between the

explanatory variable makes this estimator instable. Schaefer, Roi, and Wolfe (1984) proposed the following logistic ridge estimators to cure this problem:

$$\hat{\beta}_k = \left(\mathbf{X}'\mathbf{W}\mathbf{X} + k\mathbf{I} \right)^{-1} \mathbf{X}'\mathbf{W}\mathbf{X}\hat{\beta}_{\text{MLE}}, \quad (3)$$

where

$$k = \frac{1}{\hat{\beta}'_{\text{MLE}}\hat{\beta}_{\text{MLE}}}, \frac{p}{\hat{\beta}'_{\text{MLE}}\hat{\beta}_{\text{MLE}}}, \frac{p+1}{\hat{\beta}'_{\text{MLE}}\hat{\beta}_{\text{MLE}}}$$

The following logistic Liu estimator $\hat{\beta}_d$ was defined in Månsson, Kibria, and Shukur (2012):

$$\hat{\beta}_d = \left(\mathbf{X}'\mathbf{W}\mathbf{X} + \mathbf{I} \right)^{-1} \left(\mathbf{X}'\mathbf{W}\mathbf{X} + d\mathbf{I} \right) \hat{\beta}_{\text{MLE}}, \quad (4)$$

where $0 < d < 1$. Also, Huang (2012) defined the estimator $\hat{\beta}(k, d)$ as a combination of the two different estimators given above such that

$$\hat{\beta}(k, d) = \left(\mathbf{X}'\mathbf{W}\mathbf{X} + k\mathbf{I} \right)^{-1} \left(\mathbf{X}'\mathbf{W}\mathbf{X} + kd\mathbf{I} \right) \hat{\beta}_{\text{MLE}}, \quad (5)$$

where $k > 0$, $0 < d < 1$. It was shown that: if $d = 1$, then $\hat{\beta}(k, d) = \hat{\beta}_{\text{MLE}}$; if $k = 0$, then $\hat{\beta}(k, d) = \hat{\beta}_{\text{MLE}}$; and if $k = 1$, $\hat{\beta}(k, d) = \hat{\beta}_d$. In Månsson et al. (2012), it was proved that, when $d < 1$, $\|\hat{\beta}_d\| < \|\hat{\beta}_{\text{MLE}}\|$; it was also shown that $\hat{\beta}_d$ has a better performance than $\hat{\beta}_{\text{MLE}}$ in the presence of multicollinearity.

In order to provide the explicit form of the MSE function of $\hat{\beta}(k, d)$ use the following transformations:

Let $\alpha = \mathbf{Q}'\beta$, $\mathbf{Q}'\mathbf{X}'\mathbf{W}\mathbf{X}\mathbf{Q} = \Lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{p+1})$, where $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{p+1} > 0$ such that the λ_j 's are the eigenvalues of the matrix $\mathbf{X}'\mathbf{W}\mathbf{X}$ and $\mathbf{Q}$ is the matrix whose columns are the eigenvectors of the matrix $\mathbf{X}'\mathbf{W}\mathbf{X}$. ($\mathbf{X}'\mathbf{W}\mathbf{X} = \mathbf{Q}'\Lambda\mathbf{Q}$).

The MSE function of $\hat{\beta}(k, d)$ is as follows:

$$\begin{aligned} \text{MSE}(\hat{\boldsymbol{\beta}}(k, d)) &= \sum_{j=1}^{p+1} \frac{(\lambda_j + kd)^2}{\lambda_j (\lambda_j + k)^2} + \sum_{j=1}^{p+1} \frac{k^2 (1-d)^2 \alpha_j^2}{(\lambda_j + k)^2} \\ &= f_1(k, d) + f_2(k, d) \end{aligned} \quad (6)$$

where $f_1(k, d)$ is the variance function and $f_2(k, d)$ is the squared bias. Thus choose suitable values for the parameters k and d in order to obtain a less MSE value than that of MLE.

Three theorems given in Huang (2012) are presented about the properties of the estimator $\hat{\boldsymbol{\beta}}(k, d)$:

Theorem 1. The asymptotic variance $f_1(k, d)$ and the squared bias $f_2(k, d)$ are two continuous functions of k and d ; for fixed d^* , $0 < d^* < 1$, $f_1(k, d^*)$ and $f_2(k, d^*)$ are monotonically decreasing and increasing functions of k , respectively; for a fixed $k^* > 0$, $f_1(k^*, d)$ and $f_2(k^*, d)$ are monotonically increasing and decreasing functions of k , respectively (Huang, 2012).

Theorem 2. For a fixed d^* , $0 < d^* < 1$, there exists a $k > 0$ such that $\text{MSE}(\hat{\boldsymbol{\beta}}(k, d)) < \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{MLE}})$ (Huang, 2012).

Theorem 3. If $k > 0$ and $0 < d < 1$, then $\text{MSEM}(\hat{\boldsymbol{\beta}}(k, d)) < \text{MSEM}(\hat{\boldsymbol{\beta}}_{\text{MLE}})$ if and only if $k(1-d)\mathbf{a}'[k(1+d)\mathbf{\Lambda}^{-1} + 2\mathbf{I}]^{-1}\mathbf{a} < 1$, where $\text{MSEM}(\tilde{\boldsymbol{\beta}}) = \text{E} \left[(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})' \right]$, $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{MLE}}, \hat{\boldsymbol{\beta}}(k, d)$.

These theorems show the theoretical advantage of the estimator $\hat{\boldsymbol{\beta}}(k, d)$. In Huang (2012), the author designed a simulation study considering $k = 0.1, 0.3$, and 0.5 and $d = 0.2, 0.4$, and 0.6 . However, it is anticipated that optimal value of d and some estimators of k will result in better performance.

Note that the values of the parameter d are restricted to the interval $0 < d < 1$ in the definition of $\hat{\boldsymbol{\beta}}(k, d)$ in order to obtain a sufficient condition satisfying $\text{MSE}(\hat{\boldsymbol{\beta}}(k, d)) < \text{MSE}(\hat{\boldsymbol{\beta}}_{\text{MLE}})$. However, in the simulation given in Huang (2012), the author states that MSE of $\hat{\boldsymbol{\beta}}(k, d)$ can be smaller than MSE of $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ without satisfying this sufficient condition. Thus, we expand the restriction on d and

conduct our simulation study such that the optimal parameter d_{opt} satisfies the following conditions: $-\infty < d_{\text{opt}} < \infty$, $d_{\text{opt}} \neq 0$, and $d_{\text{opt}} \neq 1$. In the following theorem, d_{opt} is presented.

Theorem 4. The optimal shrinkage parameter $-\infty < d_{\text{opt}} < \infty$ for minimizing $\text{MSE}(\hat{\beta}(k, d)) \forall k > 0$ is the following:

$$d_{\text{opt}} = \frac{\sum_{j=1}^{p+1} \frac{k\alpha_j^2 - 1}{(\lambda_j + k)^2}}{\sum_{j=1}^{p+1} \frac{k(1 + \lambda_j\alpha_j^2)}{\lambda_j(\lambda_j + k)^2}} \quad (7)$$

Proof: It is easy to find the optimal parameter d_{opt} by differentiating $\text{MSE}(\hat{\beta}(k, d))$ with respect to d and equating the derivative to zero. Solving the equation for d , we get the optimal parameter d_{opt} . ■

After choosing the optimal parameter d_{opt} for d , the parameter k must be selected. In literature, there are some estimators for the selection of k . The followings are the estimators of k that are used in the simulation study:

1. $k_1 = \frac{p+1}{\hat{\beta}'_{\text{MLE}} \hat{\beta}_{\text{MLE}}} \text{ (Schaefer et al., 1984)}$
2. $k_2 = \frac{p}{\hat{\beta}'_{\text{MLE}} \hat{\beta}_{\text{MLE}}} \text{ (Schaefer et al., 1984)}$
3. $k_3 = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\text{max}}^2}$, where $\hat{\alpha}_{\text{max}}$ is the maximum element of $\hat{\alpha}_{\text{MLE}} = \mathbf{Q}'\hat{\beta}_{\text{MLE}}$

and $\hat{\sigma}^2 = \frac{(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})}{n - p - 1} \text{ (Hoerl \& Kennard, 1970)}$

4. $k_4 = \frac{1}{\hat{\alpha}_{\text{max}}^2}$, which is a modified version of k_3 (Schaefer et al., 1984)
5. $k_5 = \frac{\hat{\sigma}^2}{\left(\prod_{j=1}^{p+1} \hat{\alpha}_j^2\right)^{\frac{1}{p+1}}}$, which is the geometric mean of $k_j = \hat{\sigma}^2 / \hat{\alpha}_j^2$
(Kibria, 2003)

6. $k_6 = \text{median} \left(\frac{\hat{\sigma}^2}{\hat{\alpha}_j^2} \right)$ (Kibria, 2003)
7. $k_7 = \max \left(\frac{\lambda_j \hat{\sigma}^2}{(n-p-1) \hat{\sigma}^2 + \lambda_j \hat{\alpha}_j^2} \right)$ (Alkhamisi, Khalaf, & Shukur, 2006)
8. $k_8 = \max \left(\frac{1}{\sqrt{\hat{\sigma}^2 / \hat{\alpha}_j^2}} \right)$ (Muniz & Kibria, 2009)

Because the α_j^2 's and σ^2 are not known in practice, the estimators $\hat{\alpha}_j^2$ and $\hat{\sigma}^2$ are used in the above formulae.

Monte Carlo Simulation

The effective factors are chosen to be the number of explanatory variables p , the sample size n , and the correlation among the explanatory variables ρ^2 . MSE and mean absolute error (MAE) are used as the criterion of judgment.

The average MSE and MAE of the estimators $\hat{\beta}_{\text{MLE}}$ and $\hat{\beta}(k, d_{\text{opt}})$ for $k = k_1, \dots, k_8$ are computed by using the following equations:

$$\text{AMSE}(\hat{\beta}) = \sum_{r=1}^{3000} \frac{(\hat{\beta}_r - \beta)' (\hat{\beta}_r - \beta)}{3000} \quad (8)$$

$$\text{MAE}(\hat{\beta}_r) = \sum_{r=1}^{3000} \frac{|\hat{\beta}_r - \beta|}{3000}, \quad (9)$$

where $\hat{\beta}_r = \hat{\beta}_{\text{MLE}}, \hat{\beta}(k, d_{\text{opt}})$ at the r^{th} step of the simulation.

Following Kibria (2003), in order to generate the explanatory variables, the following equation is used:

$$x_{ij} = (1 - \rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip} \quad (10)$$

where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, p + 1$, and the z_{ij} 's are pseudo-random numbers following the standard normal distribution.

The dependent variable is obtained by using $\text{Be}(\mathbf{P})$, where $P_i = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$ such that x_i is the i^{th} row of the design matrix $\mathbf{X}$. The parameters $\beta_1, \beta_2, \dots, \beta_{p+1}$ are chosen due to Newhouse and Oman (1971) such that $\beta' \beta = 1$, which is a commonly used restriction in many simulation studies in the field; for example see Alkhamisi and Shukur (2008), Asar, Karaibrahimoğlu, and Genç (2014), and Kibria (2003).

The following cases are considered: $\rho^2 = 0.90, 0.95$, and 0.99 , $n = 50, 100$, and 200 , and $p = 4, 8$, and 12 . The simulation is repeated 3000 times for each set of (ρ^2, n, p) . Thus, via this set up, it may be determined which of the estimators $k_1, \dots, k_8$ has better performance when d_{opt} is used for different combinations of (ρ^2, n, p) .

Results and Discussion

Results of the Monte Carlo simulation are compiled in Tables 1-6. The factors affecting the MSE's of the estimators in the simulation are the degree of correlation ρ , the sample size n , and the number of explanatory variables p .

Tables 1, 3, and 5 are the AMSE values of Liu-type estimators, MLE for different values of k , and optimal shrinkage parameter d_{opt} when $p = 4, 8$, and 12 . According to these tables, when n and p are fixed, the increase in the correlation ρ causes an increase in the AMSE values of the estimators without exception. When $p = 12$, the increase in the correlation inflates the AMSE values drastically. The worst case is obtained when the sample size is low and the degree of correlation is high, namely, $n = 50$ and $\rho = 0.99$. When the sample size n is increased, fixing p and ρ , it has a positive effect on estimators; in other words, it can be seen that the AMSE values decreases for all of the estimators. Especially for MLE, there is a rapid decrease in the case of high correlation. When fixing n and ρ , if the number of explanatory variables is focused upon, it is observed that an increase in the value of p corresponds to an increase in the AMSE values for both MLE and the other estimators. When the number of explanatory variables is increased, one should also increase the sample size sufficiently in order to make stable estimations. The estimator having best performance among others is k_8 for all of the situations.

In Tables 2, 4 and 6, the MAE values are presented for the estimators for different values of k and optimal shrinkage parameter d_{opt} when $p = 4, 8$, and 12 . Similar comments apply in the case of AMSE. However, the only difference is that the MAE values are significantly smaller than AMSE values for all of the cases.

Particularly for the case $p = 12$, the MAE values are much smaller than the AMSE values. Again, k_8 is the best option if MAE is used as a performance criterion.

Table 1. The AMSE values for different k with optimal d for $p = 4$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	1.1453	0.5309	0.4556	2.4285	0.9978	0.6572	10.5534	3.0258	1.5615
k_2	1.1650	0.5440	0.4647	2.4853	1.0291	0.6742	10.7168	3.0939	1.5964
k_3	1.3392	0.6530	0.5187	3.0932	1.2907	0.7937	12.8152	4.1027	2.0750
k_4	1.2258	0.5760	0.4791	2.6704	1.1247	0.7122	11.3301	3.3589	1.7338
k_5	1.2328	0.5734	0.4712	2.7752	1.1296	0.7079	11.3754	3.4391	1.7882
k_6	1.2192	0.6070	0.4969	2.5722	1.1291	0.7425	10.1046	3.0120	1.6530
k_7	1.2911	0.6498	0.5196	2.8291	1.2536	0.7898	10.5026	3.3305	1.8326
k_8	1.0532	0.5352	0.4600	2.1244	0.8674	0.6469	9.4613	2.6145	1.3765
MLE	3.1825	1.1528	0.8809	7.5693	2.6893	1.3873	31.3891	11.2257	5.3554

Table 2. The MAE values for different k with optimal d for $p = 4$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	0.9224	0.7068	0.6605	1.1921	0.9009	0.7785	2.1122	1.3928	1.1036
k_2	0.9319	0.7155	0.6674	1.2168	0.9187	0.7895	2.1498	1.4144	1.1182
k_3	1.0211	0.7872	0.7069	1.4623	1.0578	0.8626	2.6491	1.7357	1.3122
k_4	0.9613	0.7367	0.6779	1.2913	0.9696	0.8129	2.2942	1.5019	1.1735
k_5	0.9685	0.7344	0.6712	1.3379	0.9714	0.8085	2.3356	1.5334	1.1979
k_6	0.9688	0.7583	0.6912	1.2818	0.9809	0.8331	2.1056	1.4233	1.1560
k_7	1.0045	0.7861	0.7078	1.3853	1.0441	0.8613	2.2242	1.5346	1.2311
k_8	0.8891	0.7134	0.6655	1.0748	0.8401	0.7774	1.9088	1.2657	1.0335
MLE	1.6172	1.0400	0.9183	2.3961	1.5279	1.1320	4.5867	3.0125	2.1434

Table 3. The AMSE values for different k with optimal d for $p = 8$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	3.6513	1.3095	0.6274	9.2806	1.8112	0.9489	20.6034	9.796	2.5225
k_2	3.6891	1.3368	0.6433	9.3837	1.8477	0.9729	20.8967	9.9669	2.584
k_3	4.8402	2.0404	0.9534	12.8513	2.8887	1.5463	32.2306	15.7924	4.8666
k_4	4.0841	1.5948	0.7691	10.6056	2.2816	1.1993	24.3956	11.7169	3.3359
k_5	3.9290	1.5937	0.7289	10.0599	2.1088	1.1405	21.4002	11.2511	3.1836
k_6	3.7811	1.6056	0.8181	9.0515	2.0722	1.2374	18.5028	9.2001	2.788
k_7	4.1232	1.9067	0.9507	9.7368	2.5571	1.4772	18.6701	10.38	3.6014
k_8	3.2690	1.1391	0.6451	8.1544	1.5275	0.8871	17.7195	8.0577	1.9944
MLE	10.6802	4.1897	1.6451	30.2246	6.2970	2.7891	75.5528	35.5189	10.9108

LIU-TYPE LOGISTIC ESTIMATORS

Table 4. The MAE values for different k with optimal d for $p = 8$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	1.2863	1.0359	0.7590	2.0771	1.1701	0.9270	3.1881	2.3832	1.3794
k_2	1.2997	1.0486	0.7693	2.1008	1.1847	0.9396	3.2277	2.4142	1.3991
k_3	1.7004	1.3507	0.9518	2.8578	1.5745	1.2091	4.7419	3.4694	2.066
k_4	1.4414	1.1654	0.8462	2.3822	1.3516	1.0517	3.7236	2.759	1.6368
k_5	1.4218	1.1648	0.8202	2.3309	1.2945	1.0236	3.4021	2.7069	1.6007
k_6	1.4011	1.1833	0.8773	2.1406	1.3002	1.0766	3.0293	2.3767	1.5059
k_7	1.5140	1.3062	0.9517	2.3067	1.4789	1.1836	3.0591	2.6345	1.7617
k_8	1.2016	0.9769	0.7801	1.8728	1.0800	0.9094	2.8762	2.0945	1.2177
MLE	2.6659	1.9465	1.2458	4.5869	2.3624	1.6143	7.4523	5.4173	3.1146

Table 5. The AMSE values for different k with optimal d for $p = 12$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	139.2714	4.5505	0.8339	379.1795	18.6147	1.1519	693.7565	29.1359	5.4585
k_2	139.5613	4.5923	0.8505	379.7215	18.7365	1.1746	695.7493	29.472	5.5575
k_3	148.6546	6.4653	1.5089	400.7071	24.0658	2.3208	768.2535	48.479	11.1368
k_4	144.0903	5.4530	1.1198	389.2163	20.9824	1.6189	723.2426	35.3703	7.631
k_5	134.0508	4.9199	1.0217	367.5066	19.4268	1.5013	662.6167	32.3825	6.7584
k_6	131.8453	4.8675	1.1637	362.3602	17.9196	1.6392	642.9656	26.1222	5.7143
k_7	134.5658	5.8933	1.4827	364.9559	20.5335	2.1233	643.2291	27.9867	7.7012
k_8	129.5472	3.9218	0.8216	359.3362	16.0274	1.0576	639.7481	23.8502	4.1557
MLE	351.1918	11.6188	2.6647	935.6592	43.8108	4.1312	2048.4312	98.5518	22.1795

Table 6. The MAE values for different k with optimal d for $p = 12$

ρ	0.9			0.95			0.99		
n	50	100	200	50	100	200	50	100	200
k_1	8.4860	1.5200	0.8701	12.8861	2.8799	1.0276	17.5892	3.7843	1.9859
k_2	8.5038	1.5332	0.8794	12.9122	2.9003	1.0382	17.6425	3.8241	2.0081
k_3	9.1525	2.0939	1.1998	13.9870	3.8298	1.4903	19.9062	5.9098	3.1408
k_4	8.8189	1.7987	1.0191	13.4347	3.2974	1.2276	18.4166	4.5498	2.4684
k_5	8.3426	1.6721	0.9695	12.6869	3.1359	1.1809	17.4483	4.3274	2.3113
k_6	8.2283	1.6989	1.0465	12.4874	2.9680	1.2456	16.7388	3.6399	2.1214
k_7	8.4317	1.9781	1.1911	12.6229	3.3975	1.4279	16.7299	3.9188	2.5582
k_8	8.0607	1.3933	0.8786	12.2928	2.5904	0.9990	16.5748	3.2482	1.7039
MLE	13.9272	2.8855	1.5911	21.3777	5.3458	1.9804	33.1953	8.6407	4.4523

Application

An empirical application is demonstrated by using a data set taken from the web site of Statistics Sweden¹. There are 290 municipalities in Sweden; eighty-three municipalities are considered. Those considered are the urban regions defined as the municipalities belonging to the Functional analysis regions Stockholm, Göteborg and Malmö, corresponding to the year 2012. The explanatory variables are defined as follows: X_1 is the population, X_2 is the number of unemployed people, X_3 is the number of newly constructed buildings, and X_4 is the number of bankrupt firms. A binary logistic regression model is set by using the dependent variable defined as follows: If there is an increase in the population of a municipality it is coded as 1; otherwise it is coded as 0.

It is observed from Table 7 that the bivariate correlations among the regressors are high (all greater than 0.91) and the condition number of the data given by $\kappa = \sqrt{\lambda_{\max} / \lambda_{\min}} = 38.3274$ shows that there is severe multicollinearity problem with this data.

For different values of k defined above and for the optimal shrinkage parameter d_{opt} , the MSEs of the estimators are given in Table 8. It can be seen from that table that k_8 has the best performance in the sense of MSE reduction.

Also, k_5 and k_6 perform quite well. If the coefficients, standard errors of the estimators, and the corresponding t-values given in Table 9 are considered, it is seen that k_5 and k_6 have very low standard errors when compared to the other estimators. Moreover, t-values corresponding to k_5 and k_6 are larger than the others in absolute value which further shows the superiority of k_5 and k_6 . Thus, k_5 and k_6 seem to be more practical for this data set. Finally, a graph of the MSE function versus k is provided in Figure 1. According to Figure 1, the MSE of $\hat{\beta}(k, d_{\text{opt}})$ has a decreasing tendency for the increasing values of the parameter k .

Table 7. The correlation matrix of the data used in the application

	X_1	X_2	X_3	X_4
X_1	1.0000	0.9937	0.9707	0.9514
X_2	0.9937	1.0000	0.9527	0.9222
X_3	0.9707	0.9527	1.0000	0.9765
X_4	0.9514	0.9222	0.9765	1.0000

LIU-TYPE LOGISTIC ESTIMATORS

Table 8. The MSEs of the estimators used in the application

	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	MLE
MSE	766.7641	776.6589	850.1244	813.5330	724.3047	693.2426	732.2801	687.6025	1894.307

Table 9. Coefficients, standard errors, and t-values of the data

Coefficients									
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	MLE
β_1	9.5174	9.5364	9.7763	9.6347	9.4529	9.3549	9.4654	9.1923	25.3187
β_2	-5.1346	-5.1261	-5.3010	-5.1732	-5.3543	-5.9703	-5.2813	-6.3168	-17.4101
β_3	2.3166	2.4417	3.3105	2.8941	1.7992	1.5479	1.8892	1.4059	3.8671
β_4	-5.7236	-5.8749	-6.8082	-6.3768	-4.9382	-4.0628	-5.1086	-3.9790	-10.9671

Standard errors									
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	MLE
β_1	1.2685	1.2749	1.3382	1.3035	1.2496	1.2513	1.2519	1.2550	3.4576
β_2	1.0505	1.0579	1.1134	1.0852	1.0167	0.9884	1.0235	0.9840	2.7109
β_3	0.8606	0.9019	1.1718	1.0450	0.6694	0.5164	0.7069	0.4866	1.3404
β_4	0.7752	0.8023	0.9745	0.8942	0.6404	0.5067	0.6686	0.4764	1.3124

t-values									
	k_1	k_2	k_3	k_4	k_5	k_6	k_7	k_8	MLE
β_1	7.5029	7.4804	7.3054	7.3911	7.5650	7.4761	7.5609	7.3244	7.3226
β_2	-4.8877	-4.8458	-4.7611	-4.767	-5.2665	-6.0405	-5.1599	-6.4195	-6.4222
β_3	2.6919	2.7073	2.8252	2.7694	2.6878	2.9977	2.6726	2.8891	2.8850
β_4	-7.3835	-7.3224	-6.9862	-7.1315	-7.7113	-8.0179	-7.6402	-8.3516	-8.3567

Summary and Conclusion

The benefits of Liu-type estimators in logistic regression were shown in the case of multicollinearity. The optimal shrinkage parameter d_{opt} used in the Liu-type estimator was defined in Huang (2012). The Monte Carlo experiment was used to evaluate the early proposed ridge regression parameters in logistic regression. Results show that the estimators chosen from the literature outperform MLE for all of the cases taken into consideration when the optimal shrinkage parameter d_{opt} is used. AMSE and MAE values of MLE become inflated when the correlation increases and the sample size decreases. Thus, researchers are advised to use Liu-type logistic estimators with the optimal shrinkage parameter in place of MLE in the presence of multicollinearity. The best performance from the estimator will be obtained if k_8 is used.

Figure 1. The MSE of $\hat{\beta}(k, d)$ function versus k

Footnotes

1. The data used in this article may be accessed through the following website:
http://www.scb.se/sv/_Hitta-statistik/

References

- Alkhamisi, M., Khalaf, G., & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics-Theory and Methods*, 35(11), 2005-2020. doi: 10.1080/03610920600762905
- Alkhamisi, M., & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics-Theory and Methods*, 37(4), 544-564. doi: 10.1080/03610920701469152

- Asar, Y., Karaibrahimoğlu, A., & Genç, A. (2014). Modified ridge regression parameters: A comparative Monte Carlo study. *Hacettepe Journal of Mathematics and Statistics*, 43(5), 827-841. Retrieved from <http://www.hjms.hacettepe.edu.tr/uploads/7742744a-b036-456d-8367-1a3eab73e172.pdf>
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. doi: 10.1080/00401706.1970.10488634
- Huang, J. (2012). A simulation research on a biased estimator in logistic regression model. In Z. Li, X. Li, Y. Liu, & Z. Cai (Eds.), *Computational Intelligence and Intelligent Systems* (389-395). Berlin, Germany: Springer. doi: 10.1007/978-3-642-34289-9_43
- Inan, D., & Erdogan, B. E. (2013). Liu-type logistic estimator. *Communications in Statistics-Simulation and Computation*, 42(7), 1578-1586. doi: 10.1080/03610918.2012.667480
- Kibria, B. M. G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics-Simulation and Computation*, 32(2), 419-435. doi: 10.1081/SAC-120017499
- Kibria, B. M. G., Månsson, K., & Shukur, G. (2012). Performance of some logistic ridge regression estimators. *Computational Economics*, 40(4), 401-414. doi: 10.1007/s10614-011-9275-x
- Liu, K. (1993). A new class of biased estimate in linear regression. *Communications in Statistics-Theory and Methods*, 22(2), 393-402. doi: 10.1080/03610929308831027
- Liu, K. (2003). Using Liu-type estimator to combat collinearity. *Communications in Statistics-Theory and Methods*, 32(5), 1009-1020. doi: 10.1081/STA-120019959
- Månsson, K., Kibria, B. M. G., & Shukur, G. (2012). On Liu estimators for the logit regression model. *Economic Modelling*, 29(4), 1483-1488. doi: 10.1016/j.econmod.2011.11.015
- Månsson, K., & Shukur, G. (2011). On ridge parameters in logistic regression. *Communications in Statistics-Theory and Methods*, 40(18), 3366-3381. doi: 10.1080/03610926.2010.500111
- Muniz, G., & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparisons. *Communications in Statistics-Simulation and Computation*, 38(3), 621-630. doi: 10.1080/03610910802592838

Newhouse, J. P., & Oman, S. D. (1971). An evaluation of ridge estimators: A report prepared for the United States Air Force Project Rand (P-716-PR).

Retrieved from

<http://www.rand.org/content/dam/rand/pubs/reports/2007/R716.pdf>

Schaefer, R. L., Roi, L. D., & Wolfe, R. A. (1984). A ridge logistic estimator. *Communications in Statistics-Theory and Methods*, 13(1), 99-113. doi: 10.1080/03610928408828664

Solution to the Multicollinearity Problem by Adding some Constant to the Diagonal

Hanan Duzan

Universiti Sains Islam Malaysia
Nilai, Malaysia

Nurul Sima Binti Mohamaed Shariff

Universiti Sains Islam Malaysia
Nilai, Malaysia

Ridge regression is an alternative to ordinary least-squares (OLS) regression. It is believed to be superior to least-squares regression in the presence of multicollinearity. The robustness of this method is investigated and comparison is made with the least squares method through simulation studies. Our results show that the system stabilizes in a region of k , where k is a positive quantity less than one and whose values depend on the degree of correlation between the independent variables. The results also illustrate that k is a linear function of the correlation between the independent variables.

Keywords: Linear models, multicollinearity, least squares method, ridge regression

---

Introduction

The ordinary least squares (OLS) estimator is the best linear unbiased estimator (BLUE). It can be used to investigate the linear relationships between the variables of interest. The model is $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$ and it is assumed that $\boldsymbol{\beta}$ is linear (each of its elements is a linear function of y , the dependent variable). The parameter $\boldsymbol{\beta}$ in OLS has the properties of being (i) unbiased where $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ is the expected value of the slope estimates of $\boldsymbol{\beta}$, which is the true $\boldsymbol{\beta}$; and (ii) consistent, where the estimator produces the minimum variance. The OLS method has some attractive statistical properties under the following assumptions:

- i. $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, where $\boldsymbol{\varepsilon}$ and $\mathbf{0}$ are $(n \times 1)$ column vectors
- ii. $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the $(n \times n)$ identity matrix
- iii. The $(n \times p)$ matrix $\mathbf{X}$ is non-stochastic
- iv. The rank of $\mathbf{X}$ is equal to the number of columns, C , in $\mathbf{X}$, and C is less than the number of observations, n

---

*Hanan Duzan is Faculty of Science and Technology. Email at: hananduzan@yahoo.com.
Dr. Shariff is Faculty of Science and Technology..*

Multicollinearity is very high-strength correlations, corresponding to singularity, among the independent variables. This phenomenon commonly occurs when a large number of independent variables are incorporated in a regression model. High-strength correlations are encountered when measuring similar dimensions and/or concepts of a phenomenon. Multicollinearity is not the only violation of the OLS assumptions. However, an accurate multicollinearity violates the assumption that the matrix $\mathbf{X}$ is given the highest rank, which makes the OLS impossible. When a model does not reach the peak, which is the inverse of $\mathbf{X}$ that cannot be defined, an infinite number of least squares solutions is obtained.

Multicollinearity has several manifestations, including: (a) small changes in the data can produce wide swings in the parameter estimates; (b) coefficients can have high standard errors and low significance even though they may be jointly significant and the coefficient of determination, R^2 , for the regression can be quite high; and (c) coefficients may have the wrong sign or implausible magnitude (Greene, 2000, p. 256). Multicollinearity increases the standard error of the coefficients, and the increased error means that the coefficient for the particular independent variable may not be close to 0. On the other hand, a multicollinearity with a low standard error can give a significant coefficient and the researcher may not come to a conclusion with null findings.

In summary, the multicollinearity misleadingly inflates the standard error in an excessive amount. In such case, the coefficient may provide high estimates of changes in the multiple regressions when only low changes can be seen in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, but it only affects the calculations related to an individual predictor. A multiple regression model with correlated predictors indicates good combination of the entire bundle of predictors which estimate the outcome variable. However, reliable results cannot be based on an individual predictor or on a set of predictors that are redundant. A high degree of multicollinearity can prevent the computer system from performing a matrix inversion while computing the regression coefficient, or it can result in an inaccurate inversion. It is noted that in discussions of the assumptions underlying regression analyses such as OLS the phrase 'no multicollinearity' is used sometimes to refer to absence of perfect multicollinearity, which is an expression of accurate (non-stochastic) linear relations among the regression model predictors.

Ridge regression is a technique for analyzing multiple regression models that may be exposed to the multicollinearity problem. The OLS regression technique provides unbiased estimates, but their variances are so large that they can be far from the actual value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors; the net effect can be highly reliable estimates

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

of the target parameters. There is a number of common biased regression techniques, the most popular of which is ridge regression. The actual definition of ridge regression is the existence of accurate linear relationships between the variables of a regression model which we can notice. In order to identify the main predictors, it is extremely vital to deal with multicollinearity where the impact is great and the interpretation, the amendments, and the analysis occur in all the linear models. The main purpose of this study is to discuss the shortcomings of OLS regression when estimating the regression coefficients in the presence of multicollinearity, and to present the ridge estimator family as an alternative to the OLS procedure.

Several authors have suggested various estimation methods to reduce the biasness problem. When Hoerl and Kennard (1970a) developed the ridge regression technique, they suggested that this method, which is also referred to as the ridge trace, can be used to solve the biasness problem. This ridge trace is a plot which illustrates the ridge regression coefficients as the main function of k . By using this ridge trace, the analyst may give a value to k at which the regression coefficients can be stabilized. Often, the regression coefficients are varied widely to get a small value of k and then they are stabilized. Choosing the smallest possible value of k (which introduces the smallest bias) ensures that the regression coefficients can remain stable. It is noted that the increasing value of k will finally drive the regression coefficients to zero. Most of the later efforts in this area have concentrated on estimating the value of the ridge parameter k . Many different techniques for estimating k have been proposed by different researchers (e.g., Hoerl & Kennard, 1970a; b; Hoerl, Kennard, & Baldwin, 1975; McDonald & Galarneau, 1975; Lawless & Wang, 1976; Dempster, Laird, & Rubin, 1977; Khalaf & Shukur, 2005; Alkhamisi, Khalaf, & Shukur, 2006; Alkhamisi & Shukur, 2008; Muniz & Kibria, 2009; Dorugade & Kashid, 2010; Jensen & Ramirez, 2012). This study investigates the shortcomings of using the OLS estimators in the presence of multicollinearity with ridge regression presented as an alternative approach. The properties of ridge regression are discussed in detail and are based on the results obtained by El-Dereny and Rashwan (2011), who have argued that this method is superior to the least-squares estimator in the presence of multicollinearity.

Methodology

Least-Squares Estimation

Consider the following P -variable regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \quad (1)$$

where $\mathbf{Y}$ is an $(n \times 1)$ column vector of observations on the dependent variable y ; $\mathbf{X}$ is an $(n \times p)$ matrix giving n observations on $p - 1$ variables, X_2 to X_p , the first column of 1s representing the intercept term; $\boldsymbol{\beta}$ is a $(p \times 1)$ column vector of the unknown parameters; and $\boldsymbol{\varepsilon}$ is an $(n \times 1)$ column vector of n disturbance terms.

The least-square estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (2)$$

In model (1), the residual, $\boldsymbol{\varepsilon}$, is assumed to be identically, independently, and normally distributed with a mean of zero and constant variance.

The variance covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}\mathbf{X})^{-1} \quad (3)$$

Alternative Variant of the Model

The $\mathbf{X}$ -scaled variables are assumed such that $\mathbf{X}'\mathbf{X}$ has the form of a correlation matrix. To recognize this, consider the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip} + \varepsilon_i, i = 1, 2, \dots, p, \quad (4)$$

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2 + \dots + \beta_p \mathbf{X}_p. \quad (5)$$

By subtracting (5) from (4), we get

$$Y_i - \mathbf{Y} = \beta_1 (X_{i1} - \mathbf{X}_1) + \beta_2 (X_{i2} - \mathbf{X}_2) + \dots + \beta_p (X_{ip} - \mathbf{X}_p) + \varepsilon_i. \quad (6)$$

The variables are then standardized to

$$\frac{Y_i - \mathbf{Y}}{S_Y}, \quad \frac{X_{ij} - \mathbf{X}_j}{S_j}, j = 1, 2, \dots, p, \quad (7)$$

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

$$S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2, \quad S_j^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{ij} - \bar{X}_j)^2, \quad j=1,2,\dots,p. \quad (8)$$

Define the following simple function of the standardized variables:

$$Y_i^* = \frac{1}{\sqrt{n-1}} \frac{(Y_i - \bar{Y})}{S_Y} \quad (9)$$

$$X_{ij}^* = \frac{1}{\sqrt{n-1}} \frac{(X_{ij} - \bar{X}_j)}{S_j}, \quad j=1,2,\dots,p \quad (10)$$

Therefore, the parameterized model with the transformed variables corresponding to model (1) is given by

$$Y_i^* = \beta_1^* X_{i1}^* + \beta_2^* X_{i2}^* + \dots + \beta_p^* X_{ip}^* + \varepsilon_i^*. \quad (11)$$

Note that

$$\beta_j^* = \frac{\beta_j S_j}{S_Y}, \quad j=1,2,\dots,p.$$

Then the least squares estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{*t} \mathbf{X}^* \right)^{-1} \mathbf{X}^{*t} \mathbf{Y}^*. \quad (12)$$

The $\mathbf{x}^*$ matrix in the model can be written as follows:

$$\mathbf{X}^* = \begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1p}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2p}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}^* & x_{n2}^* & \cdots & x_{np}^* \end{bmatrix} \quad (13)$$

so that

$$(\mathbf{X}^{*'} \mathbf{X}^*) = \begin{bmatrix} \sum_{i=1}^n x_{i1}^{*2} & \sum_{i=1}^n x_{i1}^* x_{i2}^* & \cdots & \sum_{i=1}^n x_{i1}^* x_{ip}^* \\ \sum_{i=1}^n x_{i1}^* x_{i2}^* & \sum_{i=1}^n x_{i2}^{*2} & \cdots & \sum_{i=1}^n x_{i2}^* x_{ip}^* \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{i1}^* x_{ip}^* & \sum_{i=1}^n x_{i2}^* x_{ip}^* & \cdots & \sum_{i=1}^n x_{ip}^{*2} \end{bmatrix} \quad (14)$$

Since

$$\sum_{i=1}^n x_{ij}^{*2} = \sum_{i=1}^n \left(\frac{x_{ij} - \mathbf{x}_j}{S_j \sqrt{n-1}} \right)^2 = \frac{\sum_{i=1}^n (x_{ij} - \mathbf{x}_j)^2}{n-1} \div \mathbf{S}_j^2 = 1, j = 1, 2, \dots, p \quad (15)$$

and

$$\begin{aligned} \sum_{i=1}^n x_{ij}^* x_{ik}^* &= \sum_{i=1}^n \left(\frac{x_{ij} - \mathbf{x}_j}{S_j \sqrt{n-1}} \right) \left(\frac{x_{ik} - \mathbf{x}_k}{S_k \sqrt{n-1}} \right) \\ &= \sum_{i=1}^n \frac{(x_{ij} - \mathbf{x}_j)(x_{ik} - \mathbf{x}_k)}{(n-1)S_j S_k} \\ &= \frac{\sum_{i=1}^n (x_{ij} - \mathbf{x}_j)(x_{ik} - \mathbf{x}_k)}{\sqrt{\sum_{i=1}^n (x_{ij} - \mathbf{x}_j)^2} \sqrt{\sum_{i=1}^n (x_{ik} - \mathbf{x}_k)^2}} = r_{jk}, j, k = 1, 2, \dots, p, j \neq k \end{aligned} \quad (16)$$

where r_{jk} is the simple correlation coefficient between $\mathbf{X}_j$ and $\mathbf{X}_k$, then the matrix for the transformed variables can be written as

$$(\mathbf{X}^{*'} \mathbf{X}^*) = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{12} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{1p} & r_{2p} & \cdots & 1 \end{bmatrix}, -1 \leq r_{ij} \leq 1, i \neq j. \quad (17)$$

When the number of independent variables is two, we have

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

$$C = \left(x^{*'} x^* \right)^{-1} = \frac{1}{1 - r_{12}^2} \begin{bmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{bmatrix}. \quad (18)$$

For three independent variables, Equation (17) is then replaced by

$$C = \left(X^{*'} X^* \right)^{-1} = \begin{bmatrix} \frac{1}{1 - R_{1(23)}^2} & \frac{-r_{12} + r_{13}r_{23}}{(1 - r_{23}^2)(1 - R_{1(23)}^2)} & \frac{-r_{13} + r_{12}r_{23}}{(1 - r_{23}^2)(1 - R_{1(23)}^2)} \\ \frac{-r_{12} + r_{13}r_{23}}{(1 - r_{23}^2)(1 - R_{1(23)}^2)} & \frac{1}{1 - R_{2(13)}^2} & \frac{-r_{23} + r_{12}r_{13}}{(1 - r_{13}^2)(1 - R_{2(13)}^2)} \\ \frac{-r_{13} + r_{12}r_{23}}{(1 - r_{23}^2)(1 - R_{1(23)}^2)} & \frac{-r_{23} + r_{12}r_{13}}{(1 - r_{13}^2)(1 - R_{2(13)}^2)} & \frac{1}{1 - R_{3(12)}^2} \end{bmatrix} \quad (19)$$

where

$$R_{1(23)}^2 = \frac{r_{12}^2 + r_{13}^2 - 2r_{12}r_{13}r_{23}}{1 - r_{23}^2}.$$

In the p -variables case, the diagonal elements of $C = (X^{*'} X^*)^{-1}$ can be written as follows:

$$c_{jj} = (1 - R_j^2)^{-1}, \quad j = 1, 2, \dots, p,$$

where R_j^2 is the coefficient of determination of the least squares regression of X_j^* on the remaining $(p - 1)$ regressor variables.

$$\text{Since } \text{Var}(\hat{\beta}_j) = \sigma^2 C_{jj},$$

$$R_j^2 \rightarrow 1, \quad \text{Var}(\hat{\beta}_j) \rightarrow \infty$$

Properties of the Ridge Solution

The main properties of the ridge solution are:

- i. The length of $\hat{\beta}^*$ is a decreasing function of k

- ii. The residual sum of squares is a monotone which increases as a function of k
- iii. The ridge estimator, $\hat{\beta}^*$, is a linear transformation of the least squares estimator $\hat{\beta}$

$$\hat{\beta}^* = (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-1}(\mathbf{X}^{*'}\mathbf{X}^*)\hat{\beta} \quad (20)$$

- iv. $\hat{\beta}^*$ is a biased estimator of β .

$$E(\hat{\beta}^*) = (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-1}(\mathbf{X}^{*'}\mathbf{X}^*)\beta \neq \beta \quad (21)$$

- v. The covariance of $\hat{\beta}^*$, $k > 0$, is given by

$$\text{cov}(\hat{\beta}^*) = \sigma^2 (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-1} (\mathbf{X}^{*'}\mathbf{X}^*) (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-1} \quad (22)$$

- vi. The mean square error (MSE) of $\hat{\beta}^*$ is given by

$$\text{MSE}(\hat{\beta}^*) = E\left[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)\right] \quad (23)$$

$$= \sigma^2 \sum_{i=1}^2 \frac{\lambda_i}{(\lambda_i + k)} + k^2 \beta' (\mathbf{X}^{*'}\mathbf{X}^* + k\mathbf{I})^{-2} \beta \quad (24)$$

$$= \sum_{i=1}^2 \text{var}(\hat{\beta}_i^*) + \sum_{i=1}^2 [\text{Bias}(\hat{\beta}_i^*)]^2 \quad (25)$$

where the first term on the right hand side of Equation (25) is the sum of the variance of the estimators and the second term is the sum of squared biases, which is introduced by using $\hat{\beta}^*$ rather than $\hat{\beta}$. It can be seen that the sum of variances is a decreasing function of k , while the squared bias is an increasing function of k

- vii. $\lim_{\beta \rightarrow \infty} \text{MSE}(\hat{\beta}^*) \rightarrow \infty$ and hence, for fixed k , the ridge estimator is not minimax
- viii. If $\beta'\beta$ is bounded, then there exists a $k > 0$ such that

$$\text{MSE}(\hat{\beta}^*) < \text{MSE}(\hat{\beta})$$

The Variance Inflation Factor

The variance inflation factor (VIF) can be computed using the equation

$$\text{VIF} = (1 - R_j^2)^{-1}, \quad (26)$$

where R_j is the coefficient of determination in the regression of an explanatory variable X_j on the remaining explanatory variables of the model. If X_j has a strong linear relation with other explanatory variables, then R_j^2 will be close to one and VIF values will tend to be very high. However, in the absence of any linear relations among the explanatory variables, R_j^2 will be zero and the VIF will equal one. It is known that a VIF value greater than one indicates deviation from orthogonality and has tendencies. Generally, when the $\text{VIF} > 10$, we assume that there is high multicollinearity in the data (Mardikyan & Çetin, 2008) and that the sum of squared errors (SSE) approaches 1. There always exists a $k > 0$ such that $\hat{\beta}(k)$ has smaller MSE than $\hat{\beta}$, which means that $\text{MSE}(\hat{\beta}(k)) < \text{MSE}(\hat{\beta})$. Further details on this issue have been provided by Judge (1988), Gujarati (1995), Gruber (1998), and Pasha and Shah (2004). Finally, if $R_j^2 < R^2$ for all j , and $R_j^2 < 0.90$ (which implies that $\text{VIF} = (1 - R_j^2)^{-1} < 10$), then there is no need to worry about existence of multicollinearity; the R^2 will be close to 1 and the VIF will be large. However, when $\text{VIF}_i > 10$, the data have collinearity problems.

Monte Carlo Design

A simulation study using 1000 samples with $n = 10$ was conducted to determine the appropriate k value for ridge regression in a p -variable regression model. The performances of the OLS and the different ridge regression estimators are evaluated and compared. Furthermore, a brief description of the factors that vary in our simulation study is discussed in this section.

In most simulation studies (e.g., Ghazi & Barimal, 2010), the MSE, VIF, and β of the proposed ridge estimators are calculated using a fairly low number of explanatory variables (two and four are the most common value of p). We will choose k which gives stable values of the estimated parameters and small VIF and

MSE values for different k values. A linear regression model with correlated independent variables is considered, and the different potential R_j values are computed.

The values of x_i , $i = 1, 2, 3$, and 4 were generated from normal distribution with $(0, 2)$. For given x_1, x_2, x_3, x_4 correlated variables ($0 \leq R_j^2 \leq 1$), the y values were generated using a set of predetermined values of parameters. However, only values of ε_i ($i = 1, 2, \dots, n$) were allowed to change randomly. The errors ε_i were generated to be $\varepsilon_i \sim \text{i.i.d } N(0, \sigma^2)$, i.e. independent and identically normally distributed with a mean of zero and variance of σ^2 . The true values of the parameters were taken to be:

$$\beta' = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4) = (0.2, 1.2, 0.8, 2.5, 1.2), \sigma^2 = 2.$$

One thousand data sets were used in each simulation study. Each data set was fitted by least squares and ridge regression estimation. The VIF and SSE for different k values and different R_j^2 were computed.

The mean of $\hat{\beta}_j^*$, the ridge estimates, VIF, and SSE are given by the following equations:

$$\begin{aligned} \overline{\hat{\beta}_j^*} &= \frac{\sum_{i=1}^{1000} (\hat{\beta}_j^*)_i}{1000}, \quad j = 1, 2, 3, 4 \\ \overline{\text{VIF}} &= \frac{\sum_{i=1}^{1000} (\text{VIF})_i}{1000}, \quad j = 1, 2, 3, 4 \\ \overline{\text{SSE}} &= \frac{\sum_{i=1}^{1000} (\text{SSE})_i}{1000}, \quad j = 1, 2, 3, 4 \end{aligned}$$

Results

The simulation results are presented in Tables 4 to 7 and Figures 1 to 12. The results show that the system stabilized for the various ranges of k values based on the observed ridge trace, VIF, and SSE for selected values of R_i^2 , $i = 1, 2, 3, 4$. For example, using Table 4 with $R_1^2 = 0.863$, $R_2^2 = 0.793$, $R_3^2 = 0.793$, $R_4^2 = 0.831$, $k = 0.63$ was chosen as a stable point solution. This gives $\hat{\beta}_1^* = 0.240571$, $\hat{\beta}_2^* = 0.237122$, $\hat{\beta}_3^* = 0.174854$, and $\hat{\beta}_4^* = 0.237988$, which are quite different

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

from the least squares estimates when $k = 0.63$. As a matter of fact, the R_j^2 values are large.

Table 1 lists the appropriate k values for ridge regression estimates $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \hat{\beta}_4^*$ for different R_j^2 values. Table 2 and Table 3 indicate that there is a relationship between k and R_j^2 , and that the appropriate model for the data in Table 1 is a multiple regression model given by

$$K = \theta_1 R_1^2 + \theta_2 R_2^2 + \theta_3 R_3^2 + \theta_4 R_4^2 + \varepsilon .$$

The developed model is

$$K = 0.174R_1^2 + 0.170R_2^2 + 0.194R_3^2 + 0.199R_4^2 + \varepsilon$$

Table 1. The appropriate k values for the different R_j^2 values or ridge regression models

k	R_1^2	R_2^2	R_3^2	R_4^2
0.18	0.348	0.073	0.320	0.124
0.20	0.356	0.126	0.161	0.345
0.23	0.079	0.219	0.436	0.427
0.29	0.295	0.376	0.301	0.300
0.32	0.423	0.483	0.420	0.378
0.38	0.726	0.426	0.480	0.478
0.39	0.608	0.495	0.228	0.732
0.42	0.798	0.871	0.083	0.631
0.46	0.927	0.367	0.441	0.907
0.47	0.565	0.724	0.592	0.711
0.50	0.656	0.918	0.518	0.884
0.51	0.830	0.822	0.503	0.666
0.52	0.503	0.666	0.830	0.822
0.55	0.356	0.954	0.917	0.846
0.56	0.956	0.967	0.539	0.766
0.58	0.939	0.954	0.703	0.782
0.59	0.874	0.884	0.853	0.861
0.60	0.960	0.976	0.986	0.976
0.63	0.863	0.793	0.793	0.831
0.66	0.940	0.783	0.847	0.932
0.72	0.951	0.957	0.703	0.972
0.71	0.912	0.968	0.830	0.967
0.76	0.970	0.955	0.973	0.974
0.78	0.903	0.999	0.994	0.999
0.81	0.984	0.999	0.995	0.999

Table 2. The estimated regression coefficients, the standard errors, and the associated t -tests ($R^2 = 99\%$)

Predictor	No Constant	Coef	SE Coef	T	P
R_1^2		0.17361	0.02473	7.02	0.000
R_2^2		0.17018	0.02489	6.84	0.000
R_3^2		0.19415	0.02377	8.17	0.000
R_4^2		0.19859	0.02843	6.99	0.000

Table 3. Analysis of variance

Source	DF	SS	MS	F	P
Regression	4	13.5421	3.3855	2053.91	0.000
Residual Error	56	0.0923	0.0016		
Total	60	13.6344			

Table 4. The simulation means of $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \hat{\beta}_4^*$ VIF and SSE when $R_1^2 = 0.863$, $R_2^2 = 0.793$, $R_3^2 = 0.793$, and $R_4^2 = 0.831$

K	$A\hat{\beta}_1^*$	$A\hat{\beta}_2^*$	$A\hat{\beta}_3^*$	$A\hat{\beta}_4^*$	AVIF ₁	AVIF ₂	AVIF ₃	AVIF ₄	ASSE
0.00	0.298655	0.341844	0.025654	0.381536	7.31459	4.83671	4.83955	5.90772	0.007532
0.59	0.243716	0.240542	0.173427	0.241269	1.12303	1.04726	1.06043	1.08299	0.159369
0.60	0.242624	0.239611	0.173476	0.241021	1.10790	1.03431	1.04713	1.06902	0.161539
0.61	0.241733	0.238099	0.175252	0.239907	1.09315	1.02168	1.03419	1.05542	0.162970
0.62	0.241761	0.237168	0.174536	0.239319	1.07885	1.00940	1.02159	1.04217	0.165100
0.63	0.240571	0.237122	0.174854	0.237988	1.06493	0.99739	1.00928	1.02925	0.167347
0.64	0.240347	0.236061	0.174719	0.237375	1.05134	0.98570	0.99730	1.01668	0.169175
0.65	0.238930	0.235864	0.174632	0.236797	1.03815	0.97429	0.98559	1.00444	0.171522
0.66	0.238230	0.235382	0.174621	0.235977	1.02527	0.96313	0.97417	0.99249	0.173358
0.67	0.237334	0.234947	0.174993	0.234938	1.01275	0.95225	0.96301	0.98083	0.175236
0.68	0.236631	0.233285	0.174871	0.235186	1.00053	0.94161	0.95214	0.96944	0.177417
0.69	0.236608	0.233036	0.173879	0.234140	0.98859	0.93123	0.94148	0.95836	0.179716

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

Table 5. The simulation means of $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \hat{\beta}_4^*$ VIF and SSE when $R_1^2 = 0.563$, $R_2^2 = 0.540$, $R_3^2 = 0.550$, and $R_4^2 = 0.530$

K	$A\hat{\beta}_1^*$	$A\hat{\beta}_2^*$	$A\hat{\beta}_3^*$	$A\hat{\beta}_4^*$	AVIF ₁	AVIF ₂	AVIF ₃	AVIF ₄	ASSE
0.00	0.397900	0.278189	0.031668	0.427235	2.28958	2.17575	2.22168	2.12754	0.012955
0.42	0.301376	0.243077	0.131435	0.316583	1.05823	1.03735	1.04590	1.02797	0.154597
0.43	0.301949	0.243416	0.130461	0.313638	1.04523	1.02492	1.03325	1.01581	0.157037
0.44	0.299678	0.241674	0.132085	0.313240	1.03254	1.01282	1.02089	1.00395	0.159689
0.45	0.296759	0.242362	0.133185	0.311539	1.02020	1.00099	1.00885	0.99237	0.162455
0.46	0.295325	0.241872	0.134068	0.309751	1.00812	0.98945	0.99712	0.98108	0.165044
0.47	0.293972	0.240496	0.135793	0.308160	0.99635	0.97821	0.98565	0.97003	0.167455
0.48	0.290813	0.240918	0.136362	0.307768	0.98488	0.96721	0.97444	0.95924	0.169584
0.49	0.292787	0.239215	0.134971	0.304933	0.97369	0.95647	0.96352	0.94872	0.173381
0.50	0.290226	0.238191	0.136188	0.304710	0.96275	0.94598	0.95284	0.93842	0.175646
0.51	0.288852	0.237685	0.138675	0.302297	0.95204	0.93573	0.94243	0.92835	0.177022
0.52	0.289634	0.238207	0.134940	0.300539	0.94160	0.92570	0.93221	0.91852	0.180999

Table 6. The simulation means of $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \hat{\beta}_4^*$ VIF and SSE when $R_1^2 = 0.295$, $R_2^2 = 0.376$, $R_3^2 = 0.301$, and $R_4^2 = 0.300$

K	$A\hat{\beta}_1^*$	$A\hat{\beta}_2^*$	$A\hat{\beta}_3^*$	$A\hat{\beta}_4^*$	AVIF ₁	AVIF ₂	AVIF ₃	AVIF ₄	ASSE
0.00	0.354471	0.337643	0.025927	0.522447	1.41774	1.60351	1.43014	1.42781	0.014170
0.25	0.309537	0.306022	0.079025	0.427219	1.01056	1.08747	1.01416	1.01374	0.125967
0.26	0.311451	0.300830	0.083100	0.423240	0.99922	1.07392	1.00270	1.00232	0.129788
0.27	0.306251	0.302779	0.080909	0.422513	0.98817	1.06073	0.99152	0.99115	0.134086
0.28	0.306893	0.300791	0.084654	0.417579	0.97737	1.04786	0.98059	0.98023	0.137238
0.29	0.303219	0.299035	0.085051	0.417393	0.96684	1.03533	0.96991	0.96959	0.141547
0.30	0.303990	0.295907	0.087936	0.413816	0.95650	1.02312	0.95947	0.95916	0.144839
0.31	0.299973	0.296581	0.088638	0.411886	0.94641	1.01117	0.94927	0.94897	0.148536
0.32	0.301425	0.297556	0.087440	0.406633	0.93654	0.99956	0.93927	0.93901	0.151623
0.33	0.299633	0.295511	0.088587	0.404863	0.92688	0.98817	0.92951	0.92926	0.155481
0.34	0.298779	0.293971	0.089124	0.402402	0.91742	0.97711	0.91997	0.91974	0.159396
0.35	0.295223	0.290853	0.089966	0.403828	0.90816	0.96626	0.91063	0.91040	0.162995

Table 7. The simulation means of $\hat{\beta}_1^*, \hat{\beta}_2^*, \hat{\beta}_3^*, \hat{\beta}_4^*$ VIF and SSE when $R_1^2 = 0.927$, $R_2^2 = 0.367$, $R_3^2 = 0.441$, and $R_4^2 = 0.907$

K	$A\hat{\beta}_1^*$	$A\hat{\beta}_2^*$	$A\hat{\beta}_3^*$	$A\hat{\beta}_4^*$	AVIF ₁	AVIF ₂	AVIF ₃	AVIF ₄	ASSE
0.00	0.570594	0.431284	0.044372	0.080629	13.71980	1.57886	1.78922	10.70640	0.023146
0.42	0.299998	0.335019	0.124706	0.235173	1.42950	0.86316	0.91980	1.28940	0.174125
0.43	0.298707	0.335742	0.122633	0.234379	1.40230	0.85500	0.91033	1.26700	0.176739
0.44	0.297638	0.331644	0.123887	0.235597	1.37620	0.84700	0.90106	1.24540	0.179077
0.45	0.295425	0.328881	0.127117	0.235160	1.35110	0.83917	0.89198	1.22460	0.181398
0.46	0.295129	0.326968	0.126581	0.234254	1.32700	0.83148	0.88312	1.20450	0.184709
0.47	0.292531	0.326450	0.128946	0.232747	1.30380	0.82396	0.87443	1.18520	0.186945
0.48	0.294073	0.322601	0.126818	0.233503	1.28150	0.81654	0.86591	1.16660	0.189423
0.49	0.290583	0.324226	0.125192	0.232659	1.26000	0.80928	0.85756	1.14860	0.193995
0.50	0.290984	0.320314	0.125871	0.232666	1.23930	0.80217	0.84941	1.13120	0.195867
0.51	0.288384	0.319893	0.127771	0.232432	1.21920	0.79517	0.84138	1.11440	0.197233
0.52	0.288367	0.318794	0.125241	0.231326	1.19990	0.78831	0.83354	1.09820	0.201373

Figure 1. A plot of k vs $A\hat{\beta}_1^*, A\hat{\beta}_2^*, A\hat{\beta}_3^*, A\hat{\beta}_4^*$ by using the results of a simulation study when $R_1^2 = 0.863$, $R_2^2 = 0.793$, $R_3^2 = 0.793$, and $R_4^2 = 0.831$

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

Figure 2. A plot of k vs $AVIF_1$, $AVIF_2$, $AVIF_3$, and $AVIF_4$ by using the results of a simulation study when $R_1^2 = 0.863$, $R_2^2 = 0.793$, $R_3^2 = 0.793$, and $R_4^2 = 0.831$

Figure 3. A plot of k vs SSE by using the results of a simulation study when $R_1^2 = 0.863$, $R_2^2 = 0.793$, $R_3^2 = 0.793$, and $R_4^2 = 0.831$

Figure 4. A plot of k vs $A\hat{\beta}_1^*, A\hat{\beta}_2^*, A\hat{\beta}_3^*, A\hat{\beta}_4^*$ by using the results of a simulation study when $R_1^2 = 0.563$, $R_2^2 = 0.540$, $R_3^2 = 0.550$, and $R_4^2 = 0.530$

Figure 5. A plot of k vs $AVIF_1, AVIF_2, AVIF_3$, and $AVIF_4$ by using the results of a simulation study when $R_1^2 = 0.563$, $R_2^2 = 0.540$, $R_3^2 = 0.550$, and $R_4^2 = 0.530$

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

Figure 6. A plot of k vs SSE by using the results of a simulation study when $R_1^2 = 0.563$, $R_2^2 = 0.540$, $R_3^2 = 0.550$, and $R_4^2 = 0.530$

Figure 7. A plot of k vs $A\hat{\beta}_1^*$, $A\hat{\beta}_2^*$, $A\hat{\beta}_3^*$, $A\hat{\beta}_4^*$ by using the results of a simulation study when $R_1^2 = 0.295$, $R_2^2 = 0.376$, $R_3^2 = 0.301$, and $R_4^2 = 0.300$

Figure 8. A plot of k vs $AVIF_1$, $AVIF_2$, $AVIF_3$, and $AVIF_4$ by using the results of a simulation study when $R_1^2 = 0.295$, $R_2^2 = 0.376$, $R_3^2 = 0.301$, and $R_4^2 = 0.300$

Figure 9. A plot of k vs SSE by using the results of a simulation study when $R_1^2 = 0.295$, $R_2^2 = 0.376$, $R_3^2 = 0.301$, and $R_4^2 = 0.300$

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

Figure 10. A plot of k vs $A\hat{\beta}_1^*$, $A\hat{\beta}_2^*$, $A\hat{\beta}_3^*$, $A\hat{\beta}_4^*$ by using the results of a simulation study when $R_1^2 = 0.927$, $R_2^2 = 0.376$, $R_3^2 = 0.441$, and $R_4^2 = 0.907$

Figure 11. A plot of k vs $AVIF_1$, $AVIF_2$, $AVIF_3$, and $AVIF_4$ by using the results of a simulation study when $R_1^2 = 0.927$, $R_2^2 = 0.376$, $R_3^2 = 0.441$, and $R_4^2 = 0.907$

Figure 12. A plot of k vs SSE by using the results of a simulation study when $R_1^2 = 0.927$, $R_2^2 = 0.376$, $R_3^2 = 0.441$, and $R_4^2 = 0.907$

Discussion and Conclusion

The main goal of this study was to identify the most relevant k value for ridge regression in a four-variable regression model. Since it is not possible to achieve this mathematically, a simulation study was conducted to study the behavior of ridge regression in such case. It was assumed that both the form of the model and the nature of the errors, ε_i ($i = 1, 2, \dots, n$), are known. For given X_1, X_2, X_3 , and X_4 correlated values ($0 \leq R_j^2 \leq 1$), the y values were generated using a set of predetermined values of parameters, allowing only the values of ε_i to change randomly. The errors, ε_i , were generated such that $\varepsilon_i \sim \text{i.i.d } N(0, \sigma^2)$. One thousand random data sets were used in each simulation study. The p -variable linear regression model was fit by the least-squares method. Thereupon, in each simulation study, one thousand ridge regression estimates of $\hat{\beta}^*$, VIF, and SSE for different k and R_j^2 values were computed. The simulation outcomes illustrate that there is a statistically-significant relationship between k and R_j^2 . The most appropriate model to describe the relation between k and R_j^2 is a multiple regression model and the fitted regression equation is

SOLUTION TO THE MULTICOLLINEARITY PROBLEM USING RR

$$k = 0.174R_1^2 + 0.170R_2^2 + 0.194R_3^2 + 0.199R_4^2 \quad (27)$$

From a practical point of view, when multicollinearity occurs, we can use ridge regression to solve this problem. The appropriate value of k can be chosen according to the ridge trace and Equation (27). In this study we have only considered four independent variables.

References

- Alkhamisi, M., Khalaf, G., & Shukur, G. (2006). Some modifications for choosing ridge parameters. *Communications in Statistics – Theory and Methods*, 35(11), 2005-2020. doi: [10.1080/03610920600762905](https://doi.org/10.1080/03610920600762905)
- Alkhamisi, M. A., & Shukur, G. (2008). Developing ridge parameters for SUR model. *Communications in Statistics – Theory and Methods*, 37(4), 544-564. doi: [10.1080/03610920701469152](https://doi.org/10.1080/03610920701469152)
- Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1), 1-38. Available from <http://www.jstor.org/stable/2984875>
- Dorugade, A. V., & Kashid, D. N. (2010). Alternative method for choosing ridge parameter for regression. *Applied Mathematical Sciences*, 4(9), 447-456. Retrieved from <http://www.m-hikari.com/ams/ams-2010/ams-9-12-2010/dorugadeAMS9-12-2010.pdf>
- El-Dereny, M., & Rashwan, N. I. (2011). Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Sciences*, 6(12), 585-600. Retrieved from <http://www.m-hikari.com/ijcms-2011/9-12-2011/rashwanIJCMS9-12-2011.pdf>
- Gruber, H. J. (1998). *Improving efficiency by shrinkage*. New York, NY: Marcel Dekker.
- Gujarati, D. N. (1995). *Basic econometrics*. New York, NY: McGraw-Hill.
- Hoerl, A. E., & Kennard, R. W. (1970a). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55-67. doi: [10.1080/00401706.1970.10488634](https://doi.org/10.1080/00401706.1970.10488634)
- Hoerl, A. E., & Kennard, R. W. (1970b). Ridge regression: Applications to nonorthogonal problems. *Technometrics*, 12(1), 69-82. doi: [10.1080/00401706.1970.10488635](https://doi.org/10.1080/00401706.1970.10488635)

- Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics – Theory and Methods*, 4(2), 105-123. doi: [10.1080/03610927508827232](https://doi.org/10.1080/03610927508827232)
- Jensen, D. R., & Ramirez, D. E. (2012). Variations on ridge traces in regression. *Communications in Statistics – Simulation and Computation*, 41(2), 265-278. doi: [10.1080/03610918.2011.586482](https://doi.org/10.1080/03610918.2011.586482)
- Judge, G. G. (1988). *Introduction to theory and practice of econometrics*. New York, NY: John Wiley and Sons.
- Khalaf, G., & Shukur, G. (2005). Choosing ridge parameter for regression problems. *Communications in Statistics – Theory and Methods*, 34(5), 1177-1182. doi: [10.1081/STA-200056836](https://doi.org/10.1081/STA-200056836)
- Lawless, J. F., & Wang, P. (1976). A simulation study of ridge and other regression estimators. *Communications in Statistics – Theory and Methods*, 5(4), 307-323. doi: [10.1080/03610927608827353](https://doi.org/10.1080/03610927608827353)
- Mardikyan, S., & Çetin, E. (2008). Efficient choice of biasing constant for ridge regression. *International Journal of Contemporary Mathematical Sciences*, 3(11), 527-536. Retrieved from <http://www.m-hikari.com/ijcms-password2008/9-12-2008/cetinIJCMS9-12-2008.pdf>
- McDonald, G. C., & Galarneau, D. I. (1975). A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350), 407-416. doi: [10.1080/01621459.1975.10479882](https://doi.org/10.1080/01621459.1975.10479882)
- Muniz, G., & Kibria, B. M. G. (2009). On some ridge regression estimators: An empirical comparisons. *Communications in Statistics – Simulation and Computation*, 38(3), 621-630. doi: [10.1080/03610910802592838](https://doi.org/10.1080/03610910802592838)
- Pasha, G. R., & Shah, M. A. (2004). Application of ridge regression to multicollinear data. *Journal of Research (Science)*, 15(1), 97-106. Retrieved from <http://www.bzu.edu.pk/jrscience/vol15no1/15.php>

The xgamma Distribution: Statistical Properties and Application

Subhradev Sen
Malabar Cancer Centre
Thalassery, Kerala, India

Sudhansu S. Maiti
Visva-Bharati University
Santiniketan, India

N. Chandra
Pondicherry University
Puducherry, India

A new probability distribution, the xgamma distribution, is proposed and studied. The distribution is generated as a special finite mixture of exponential and gamma distributions and hence the name proposed. Various mathematical, structural, and survival properties of the xgamma distribution are derived, and it is found that in many cases the xgamma has more flexibility than the exponential distribution. To evaluate the comparative behavior, stochastic ordering of the distribution is studied. To estimate the model parameter, the method of moment and the method of maximum likelihood estimation are proposed. A simulation algorithm to generate random samples from the xgamma distribution is indicated along with a simulation study. A real life dataset on the remission times of patients receiving an analgesic is analyzed, and it is found that the xgamma model provides better fit to the data as compared to the exponential model.

Keywords: Finite mixture distribution, lifetime distribution, survival properties, maximum likelihood estimation

---

Introduction

The exponential and gamma are well known probability distributions used for modeling lifetime data. Both distributions possess some interesting structural properties, for example, exponential distribution possesses memory less and constant hazard rate properties. Moreover, as a special case of the gamma distribution, the exponential distribution can be used in modeling time-to-event data or modeling waiting times. Various extensions of both distributions can be obtained in the literature for describing the uncertainty behind real life phenomena arising in the area of survival modeling (see [Johnson, Kotz, & Balakrishnan, 1994; 1995; Lawless, 2002](#)) and reliability engineering (for more details, see [Barlow & Proschan, 1981](#)). The introduction of new lifetime distributions or modified lifetime distributions has become a time-honored fashion in statistical and biomedical

---

Subhradev Sen is a Lecturer in Biostatistics. Email him at: s.subhradev@gmail.com.

research. Finite mixture distributions arising from the standard distributions play, in most situations, a better role in modeling real-life phenomena as compared to the standard ones (see [Mclachlan & Peel, 2000](#)).

With the advances in technology and science, a wealth of information has allowed the statistician and data modeler to think in a broader way in the process of gathering knowledge. In order to make inferences about the population of interest, statisticians gather and analyze these information keeping, the responsibility of accurate inference. In recent years, it has been observed that many well-known distributions used to model data sets do not offer enough flexibility to provide an adequate fit. It is, therefore, the need of time that guides statisticians to model real-life scenarios by introducing distributions that are more flexible.

Keeping the role of finite mixture distributions (for more details, see [Everitt, 1996](#); [Everitt & Howell, 2005](#)) in modeling time-to-event data, a new distribution, namely the xgamma distribution, is introduced in this article. A special mixture of exponential and gamma distributions is considered in order to obtain the form of xgamma and hence the name proposed. The basic structural and survival properties of the xgamma distribution are obtained in the subsequent sections. Maximum Likelihood and method of moments estimators for the parameters of the model are found, as is a simulation study algorithm, and finally a real data application is made to show the superiority of the xgamma distribution over the exponential distribution.

Methodology & Synthesis

A special finite mixture of exponential and gamma distributions is used to obtain a new probability distribution, called the xgamma distribution. A random variable X is said to have a finite mixture distribution if its probability density function (pdf) $f(x)$ is of the form

$$f(x) = \sum_{i=1}^k \pi_i f_i(x), \quad (1)$$

where each $f_i(x)$ is a pdf and $\pi_1, \pi_2, \dots, \pi_k$ denote the mixing proportions that are non-negative and sum to one.

We have considered $f_1(x)$ to follow an exponential distribution with parameter θ and $f_2(x)$ to follow a gamma distribution with scale parameter θ and shape parameter 3, i.e. $f_1(x) \sim \text{Exp}(\theta)$ and $f_2(x) \sim \text{Gamma}(3, \theta)$ with $\pi_1 = \theta/(1 + \theta)$ and $\pi_2 = 1 - \pi_1$.

THE XGAMMA DISTRIBUTION

Definition: A continuous random variable X is said to follow an xgamma distribution if its pdf is of the form

$$f(x) = \frac{\theta^2}{(1+\theta)} \left(1 + \frac{\theta}{2} x^2 \right) e^{-\theta x}, \quad x > 0, \theta > 0 \quad (2)$$

and is denoted by $X \sim \text{xgamma}(\theta)$. The cumulative density function (cdf) of X is given by

$$F(x) = 1 - \frac{\left(1 + \theta + \theta x + \frac{\theta^2 x^2}{2} \right)}{(1+\theta)} e^{-\theta x}, \quad x > 0, \theta > 0 \quad (3)$$

Shape

The first derivative of (2) is

$$\frac{d}{dx} f(x) = \frac{\theta^2}{(1+\theta)} \left(\theta x - \theta - \frac{\theta^2}{2} x^2 \right) e^{-\theta x}.$$

It follows that

- i. For $\theta \leq \frac{1}{2}$, $\frac{d}{dx} f(x) = 0$ implies that $\frac{1 + \sqrt{1 - 2\theta}}{\theta}$ is the unique critical point at which $f(x)$ is maximized
- ii. For $\theta > \frac{1}{2}$, $\frac{d}{dx} f(x) \leq 0$, i.e. $f(x)$ is decreasing in x

Figure 1 shows the pdf of the xgamma distribution given in (2) for selected values of θ .

The mode of the xgamma distribution is given by

$$\text{Mode}(X) = \begin{cases} \frac{1 + \sqrt{1 - 2\theta}}{\theta}, & 0 < \theta < \frac{1}{2} \\ 0, & \text{otherwise} \end{cases}$$

Figure 1. Probability density of $xgamma(\theta)$ for selected values of θ

Remark: It is to be noted that the mode of the exponential distribution is always 0 while the mode of $xgamma$ can be varied as seen above. It is easy to show that if $X \sim xgamma(\theta)$, then $\text{Mode}(X) < \text{Median}(X) < \text{Mean}(X)$, which also holds good for exponential distribution.

Moments and Related Measures

The r^{th} moment about the origin of $xgamma$ distribution is

$$\mu'_r = E(X^r) = \frac{r!(\theta + r + a_r)}{\theta^r(1 + \theta)},$$

where $a_r = a_{r-1} + r$ for $r = 1, 2, 3, \dots$ with $a_0 = 0$ and $a_1 = 2$. In particular,

$$\begin{aligned} \mu'_1 &= \frac{(\theta + 3)}{\theta(1 + \theta)} = \text{Mean}(X) = \mu \text{ (say)} \\ \mu'_2 &= \frac{2(\theta + 6)}{\theta^2(1 + \theta)}, \mu'_3 = \frac{6(\theta + 10)}{\theta^3(1 + \theta)}, \mu'_4 = \frac{24(\theta + 15)}{\theta^4(1 + \theta)} \end{aligned}$$

THE XGAMMA DISTRIBUTION

It is to be noted that, for the exponential distribution with parameter θ , the r^{th} order moment about origin is

$$\mu'_r = \frac{r!}{\theta^r}.$$

The j^{th} order central moment of the xgamma distribution is $\mu_j = E[(X - \mu)^j] = \sum_{r=0}^j \binom{j}{r} \mu'_r (-\mu)^{j-r}$. In particular,

$$\begin{aligned}\mu_2 &= \frac{(\theta^2 + 8\theta + 3)}{\theta^2 (1 + \theta)^2} = \text{var}(X) = \sigma^2 \text{ (say)} \\ \mu_3 &= \frac{2(\theta^3 + 15\theta^2 + 9\theta + 3)}{\theta^3 (1 + \theta)^3} \\ \mu_4 &= \frac{3(5\theta^4 + 88\theta^3 + 310\theta^2 + 288\theta + 177)}{\theta^4 (1 + \theta)^4}\end{aligned}$$

The coefficients of variation (γ), skewness ($\sqrt{\beta_1}$), and kurtosis (β_2) are

$$\begin{aligned}\gamma &= \frac{\sqrt{\theta^2 + 8\theta + 3}}{(\theta + 3)} \\ \sqrt{\beta_1} &= \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{2(\theta^3 + 15\theta^2 + 9\theta + 3)}{(\theta^2 + 8\theta + 3)^{3/2}} \\ \beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{3(5\theta^4 + 88\theta^3 + 310\theta^2 + 288\theta + 177)}{(\theta^2 + 8\theta + 3)^2}\end{aligned}$$

The coefficients are increasing functions in θ (see [Figure 2](#) for the graph of γ and $\sqrt{\beta_1}$ for varying θ).

Figure 2. Coefficients for variation and skewness

Remark: It should be noted that the values of γ , $\sqrt{\beta_1}$, and β_2 for the exponential distribution are 1, 2, and 6, respectively. Hence the xgamma distribution is again more flexible than the exponential distribution.

Survival Properties

The hazard rate function or failure rate function for a continuous distribution with pdf $f(x)$, cdf $F(x)$, and survival function (sf) $S(x)$ is defined as

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{P(X < x + \Delta x | X > x)}{\Delta x} = \frac{f(x)}{1 - F(x)} = \frac{f(x)}{S(x)}.$$

For the xgamma distribution, the hazard rate function is given by

THE XGAMMA DISTRIBUTION

$$h(x) = \frac{\theta^2 \left(1 + \frac{\theta}{2} x^2\right)}{\left(1 + \theta + \theta x + \frac{\theta^2 x^2}{2}\right)} \quad (4)$$

It is to be noted that

- i. $h(0) = \frac{\theta^2}{(1+\theta)} = f(0)$
- ii. $h(x)$ is an increasing function in x and θ with $\theta^2/(1+\theta) < h(x) < \theta$

Remark: For the exponential distribution with parameter θ , $h(x) = \theta$, and so (4) shows the flexibility of the xgamma distribution over the exponential distribution. Figure 3 shows the hazard rate function of the xgamma distribution for selected values of θ .

Figure 3. Hazard rate function of xgamma(θ) for selected values of θ

Figure 4. Mean residual life function plot of $x\gamma(\theta)$ for selected values of θ

For a continuous random variable X with pdf $f(x)$ and cdf $F(x)$, the mean residual life (mrl) function is defined as

$$m(x) = E(X - x | X > x) = \frac{1}{1 - F(x)} \int_x^{\infty} [1 - F(t)] dt.$$

For the $x\gamma$ distribution, the mrl function (see Figure 4) is given by

$$\begin{aligned} m(x) &= \frac{1}{\left(1 + \theta + \theta x + \frac{\theta^2 x^2}{2}\right) e^{-\theta x}} \int_x^{\infty} \left(1 + \theta + \theta t + \frac{\theta^2 t^2}{2}\right) e^{-\theta t} dt \\ &= \frac{1}{\theta} + \frac{(x + \theta x)}{\theta \left(1 + \theta + \theta x + \frac{\theta^2 x^2}{2}\right)} \end{aligned} \quad (5)$$

It is to be noted that

- i. $m(0) = \mu = \frac{(\theta+3)}{\theta(1+\theta)}$
- ii. $m(x)$ is decreasing in x and θ with $\frac{1}{\theta} < m(x) < \frac{(\theta+3)}{\theta(1+\theta)}$

Remark: For the exponential distribution, the mrl function is $1/\theta$ and hence (6) again shows the flexibility of the xgamma distribution over the exponential distribution.

Stochastic Ordering

For a positive continuous random variable, stochastic ordering is an important tool for judging the comparative behavior. Recall some basic definitions:

A random variable X is said to be smaller than a random variable Y in the

- i. stochastic order ($X \leq_{st} Y$) if $F_X(x) \geq F_Y(x)$ for all x
- ii. hazard rate order ($X \leq_{hr} Y$) if $h_X(x) \geq h_Y(x)$ for all x
- iii. mean residual life order ($X \leq_{mrl} Y$) if $m_X(x) \leq m_Y(x)$ for all x
- iv. likelihood ratio order ($X \leq_{lr} Y$) if $f_X(x)/f_Y(x)$ decreases in x

The following implications (see Shaked & Shanthikumar, 1994) are well justified:

$$\begin{aligned} X \leq_{lr} Y &\Rightarrow X \leq_{hr} Y \Rightarrow X \leq_{mrl} Y, \\ X \leq_{hr} Y &\Rightarrow X \leq_{st} Y \end{aligned} \quad (6)$$

The following theorem shows that the xgamma distributions are ordered with respect to the strongest likelihood ratio ordering.

Theorem 1. Let $X \sim \text{xgamma}(\theta_1)$ and $Y \sim \text{xgamma}(\theta_2)$. If $\theta_1 > \theta_2$ then $X \leq_{lr} Y$ and hence the other ordering in (7).

Proof: Note that

$$\frac{f_X(x)}{f_Y(x)} = \frac{\theta_1^2 (1+\theta_2) (2+\theta_1 x^2)}{\theta_2^2 (1+\theta_1) (2+\theta_2 x^2)} e^{(\theta_2 - \theta_1)x}, \quad x > 0$$

$$\frac{d}{dx} \left(\frac{f_x(x)}{f_y(x)} \right) = (\theta_2 - \theta_1) \frac{\theta_1^2 (1 + \theta_2)}{\theta_2^2 (1 + \theta_1)} e^{(\theta_2 - \theta_1)x} \left[\frac{(2 + \theta_1 x^2)}{(2 + \theta_2 x^2)} - \frac{4x}{(2 + \theta_2 x^2)^2} \right] < 0$$

since $\theta_1 > \theta_2$. Hence $f_x(x)/f_y(x)$ decreases in x and $X \leq_{lr} Y$. The remaining statement follows from (7) directly.

Estimation of the Parameter

Given a random sample $X_1, X_2, \dots, X_n$ of size n from the xgamma distribution in (2), the method of moment (mom) estimator for the parameter θ of xgamma distribution given in (2) is obtained as follows:

Equate sample mean $\bar{X} = \sum_{i=1}^n X_i / n$ with first order moment about origin of (2) which gives the mom estimator of θ as

$$\hat{\theta}_{\text{mom}} = \frac{-(\bar{X} - 1) + \sqrt{(\bar{X} - 1)^2 + 12\bar{X}}}{2\bar{X}}, \quad \bar{X} > 0.$$

The following theorem shows that the mom estimator of θ in (2) is positively biased:

Theorem 2. The method of moment estimator of the xgamma distribution is positively biased, i.e., $E(\hat{\theta}_{\text{mom}}) - \theta > 0$.

Proof: Let $\hat{\theta}_{\text{mom}} = g(\bar{X})$ and $g(t) = \frac{-(t-1) + \sqrt{(t-1)^2 + 12t}}{2t}$. For $t > 0$, $g''(t) > 0$ and hence $g(t)$ is strictly convex. Thus by Jensen's inequality, we have $g[E(\bar{X})] < E[g(\bar{X})]$. Now since $g[E(\bar{X})] = g(\mu) = g((\theta+3)/(\theta(1+\theta))) = \theta$, we obtain $E(\hat{\theta}_{\text{mom}}) - \theta > 0$. Hence the proof.

It should be noted that the sample raw moments are unbiased and consistent estimators of the corresponding population raw moments. They are also

THE XGAMMA DISTRIBUTION

asymptotically normally distributed (CAN estimators) by virtue of the central limit theorem. Thus the mom estimator, $\hat{\theta}_{\text{mom}}$, of θ for xgamma distribution is consistent.

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)$ be n observations on a random sample $X_1, X_2, \dots, X_n$ of size n drawn from the xgamma distribution in (2). The maximum likelihood estimator (mle) of θ for given $\mathbf{x}$ is as follows:

The likelihood function is given by

$$L(\theta | x_1, x_2, \dots, x_n) = L(\theta | x) = \prod_{i=1}^n \frac{\theta^2}{(1+\theta)} \left(1 + \frac{\theta}{2} x_i^2\right) e^{-\theta x_i} .$$

The log-likelihood function is given by

$$\log_e L(\theta | \mathbf{x}) = 2n \log \theta - n \log(1+\theta) + \sum_{i=1}^n \log \left(1 + \frac{\theta}{2} x_i^2\right) - \theta \left(\sum_{i=1}^n x_i\right) . \quad (7)$$

The log-likelihood equation corresponding to (7) becomes

$$l(\theta | \mathbf{x}) = \frac{d}{d\theta} \log_e L(\theta | x) = 0 ,$$

where $l(\theta | \mathbf{x})$ is given by

$$\sum_{i=1}^n \frac{\frac{x_i^2}{2}}{\left(1 + \frac{\theta}{2} x_i^2\right)} + \frac{2n}{\theta} - \frac{n}{(1+\theta)} - \sum_{i=1}^n x_i$$

To obtain the mle of θ , $\hat{\theta}_{\text{mle}}$ (say), we can maximize (8) directly with respect to θ or we can solve the non-linear equation $l(\theta | \mathbf{x}) = 0$. Note that $\hat{\theta}_{\text{mle}}$ cannot be solved analytically; numerical iteration techniques, such as the Newton-Raphson algorithm, are thus adopted to solve the log-likelihood equation for which (8) is maximized. The initial solution for such an iteration can be taken as:

$$\theta_0 = \frac{n}{\sum_{i=1}^n x_i} .$$

Using this initial solution, we have

$$\theta^{(i)} = \theta^{(i-1)} - \frac{\mathbf{1}(\theta^{(i-1)} | \mathbf{x})}{\mathbf{1}'(\theta^{(i-1)} | \mathbf{x})}$$

for the i^{th} iteration. We choose $\theta^{(i)}$ such that $\theta^{(i)} \cong \theta^{(i-1)}$.

Remark. The method of moment and maximum likelihood estimators of the exponential distribution is $1/\bar{X}$, which is also biased and consistent.

Simulation Study

The inversion method for generating random data from the xgamma distribution fails because the equation $F(x) = u$, where u is an observation from the uniform distribution on $(0, 1)$, cannot be explicitly solved in x . However, we can use the fact that the xgamma distribution is a special mixture of the exponential(θ) and gamma(3, θ) distributions.

To generate random data $X_i, i = 1, 2, \dots, n$, from the xgamma distribution with parameter θ , we can use the following algorithm:

1. Generate $U_i \sim \text{uniform}(0, 1), i = 1, 2, \dots, n$
2. Generate $V_i \sim \text{exponential}(\theta), i = 1, 2, \dots, n$
3. Generate $W_i \sim \text{gamma}(3, \theta), i = 1, 2, \dots, n$
4. If $U_i \leq \theta/(1 + \theta)$, then set $X_i = V_i$. Otherwise, set $X_i = W_i$

A Monte Carlo simulation study was carried out considering $N = 10000$ times for selected values of n and θ . Samples of sizes 20, 40, and 100 were considered and values of θ were taken as 0.1, 0.5, 1.0, 1.5, 3, and 6. The following two measures were computed:

- i. Average bias of the simulated estimates $\hat{\theta}_i, i = 1, 2, \dots, N$:

$$\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta)$$
- ii. Average Mean Square Error (MSE) of the simulated estimates $\hat{\theta}_i, i = 1, 2, \dots, N$:

$$\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta)^2$$

THE XGAMMA DISTRIBUTION

Table 1. Average bias and MSE of the estimator $\hat{\theta}$

n theta	20		40		100	
	Bias	MSE	Bias	MSE	Bias	MSE
0.1	0.00193	0.00020	0.00078	0.00009	0.00034	0.00004
	-0.06375	0.00409	-0.06420	0.00414	-0.06438	0.00415
0.5	0.01182	0.00595	0.00539	0.00275	0.00199	0.00106
	-0.27887	0.07934	-0.28258	0.08057	-0.28455	0.08125
1.0	0.02655	0.02750	0.01347	0.01290	0.00517	0.00499
	-0.48093	0.24223	-0.49040	0.24558	-0.49631	0.24828
1.5	0.04411	0.07234	0.02804	0.03395	0.00871	0.01221
	-0.63116	0.43490	-0.64478	0.43260	-0.65975	0.44133
3.0	0.12181	0.36497	0.05765	0.15694	0.02204	0.05985
	-0.88938	1.04281	-0.94784	1.00708	-0.98001	1.00190
6.0	0.27864	1.75155	0.14144	0.78423	0.05511	0.28684
	-1.06299	2.59379	-1.19641	2.09340	-1.28001	1.88262

The result of the simulation study has been tabulated in Table 1. In Table 1, for each selected value of θ , the corresponding values relating to the xgamma distribution have been presented in first row and that relating to exponential distribution in second row.

Remarks. i) Table 1 shows that the bias is positive in the case of the xgamma distribution (as shown in the Theorem 2). Table 1 also shows that bias and MSE decreases as n increases and increases when θ increases. ii) In terms of bias and MSE of the estimates of θ , the xgamma distribution shows more flexibility as compared to the exponential distribution.

Application

In this section, a real data set is used to show that the xgamma distribution can be a better model than one based on the exponential distribution. The data on relief times (in hours) of 20 patients receiving an analgesic (cf. Gross & Clark, 1975) is used. Both the xgamma and exponential distributions are fitted to this data set. The method of maximum likelihood is used. Maximum likelihood estimates for both the cases are calculated for the data. The required numerical evaluations are carried out using R 3.1.1 software. Table 2 provides the maximum likelihood estimates with corresponding standard errors of the model parameters. The model selection is carried out using the log-likelihood value, AIC (Akaike information criterion), the AICc (consistent Akaike information criteria) and the BIC (Bayesian information criterion):

$$AIC = -2 \log_e L + 2k, \quad AICc = AIC + \frac{2k(k+1)}{n-k-1}, \quad BIC = k \log_e(n) - 2 \log_e L$$

where $\log_e L$ denotes the log-likelihood function evaluated at the maximum likelihood estimates, k is the number of parameters, and n is the sample size.

From Table 2, it is clear that the values of the AIC, AICc and BIC are smaller for the xgamma distribution compared with those values of the exponential model, so the new distribution seems to be a very competitive model to these data. It follows that the xgamma distribution provides the better fit to the data.

Table 2. Maximum likelihood estimates and model selection statistics

Distributions	Maximum Likelihood Estimates	Standard Error	Log-likelihood	AIC	AICc	BIC
Exponential	$\hat{\theta}_{mle} = 0.52632$	0.11769	-32.83708	67.67416	67.89638	68.66989
xgamma	$\hat{\theta}_{mle} = 1.10747$	0.16943	-31.50824	65.01649	65.23871	66.01221

Conclusion

The xgamma distribution, a special finite mixture of exponential and gamma distributions, was derived. Various mathematical and structural properties of the distribution were studied including the shape, moments, measures of skewness, and kurtosis. Important survival properties like the hazard rate and mean residual life functions were derived and discussed. Stochastic ordering and a simulation algorithm were also proposed. Added flexibility over the exponential distribution was observed with regard to certain important properties of the xgamma distribution. The maximum likelihood method and method of moments were proposed for the parameter estimation. In order to demonstrate the applicability of the xgamma distribution, a simulation study was shown. Moreover, the distribution was fitted to a real data set and compared with the exponential distribution. Results show that the xgamma distribution provides an adequate fit for the data set. The maximum likelihood functions may be further studied under different types of censoring mechanism for future applications of the xgamma model. Bayesian estimation of the parameter of xgamma distribution may further be considered with suitable prior and risk function.

References

- Barlow, R. E., & Proschan, F. (1981). *Statistical theory of reliability and life testing: Probability models*. Silver Spring, MD: To Begin With.
- Everitt, B. S. (1996). An introduction to finite mixture distributions. *Statistical Methods in Medical Research*, 5(2), 107-127. doi: [10.1177/096228029600500202](https://doi.org/10.1177/096228029600500202)
- Everitt, B. S., & Howell, B. (Eds.). (2005). *Encyclopedia of statistics in behavioral science*. Hoboken, NJ: John Wiley & Sons.
- Gross, A. J., & Clark, V. A. (1975). *Survival distributions: Reliability applications in the biomedical sciences*. New York, NY: John Wiley & Sons.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions* (Vol. 1) (2nd ed.). New York, NY: Wiley.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (2nd ed.) (Vol. 2). New York, NY: Wiley.
- Lawless, J. F. (2003). *Statistical models and methods for lifetime data* (2nd ed.). New York, NY: John Wiley & Sons.
- McLachlan, G., & Peel, G. (2000). *Finite mixture models*. New York, NY: John Wiley & Sons.
- Shaked, M., & Shanthikumar, J. G. (1994). *Stochastic orders and their applications*. New York, NY: Academic Press.

Model-Based Outlier Detection System with Statistical Preprocessing

D. Asir Antony Gnana Singh

Anna University
Tiruchirappalli, India

E. Jebalamar Leavline

Anna University
Tiruchirappalli, India

Reliability, lack of error, and security are important improvements to quality of service. Outlier detection is a process of detecting the erroneous parts or abnormal objects in defined populations, and can contribute to secured and error-free services. Outlier detection approaches can be categorized into four types: statistic-based, unsupervised, supervised, and semi-supervised. A model-based outlier detection system with statistical preprocessing is proposed, taking advantage of the statistical approach to preprocess training data and using unsupervised learning to construct the model. The robustness of the proposed system is evaluated using the performance evaluation metrics sum of squared error (SSE) and time to build model (TBM). The proposed system performs better for detecting outliers regardless of the application domain.

Keywords: Outlier, Preprocessing, Inter-quartile range, Anomaly Detection

---

Introduction

Outlier detection is the process of identifying the objects or events that does not follow the stranded pattern or change in behaviour compared to the group members. The outlier detection is also known as anomaly detection (Chandola, Banerjee, & Kumar, 2009). Anomaly detection helps to solve bank fraudulent detection, network or host-based intrusion detection, credit cards fraudulent detection (Akhilomen, 2013), adversaries detection in military environment, surveillance for enemy activities, detecting the outliers from captured information in ad-hoc and wireless sensor networks (Singhal, Gankotiya, Agarwal, & Verma, 2012), structural defect identification, error detection in intensive and safety systems, detecting errors in text, detecting the anomaly in computer networks to prevent the attack (Tartakovsky, Polunchenko, & Sokolov, 2013), detecting

---

D. Asir Antony Gnana Singh is in the Department of Computer Science and Engineering. Email at: asirantony@gmail.com. Dr. Leavline is a Professor in the Department of Electronics Communication and Engineering. Email her at: jebi.lee@gmail.com.

abnormal reading of radars due to the fault in the parts of the air craft (Fujimaki, Yairi, & Machida, 2005), detecting chromosomal abnormalities in medical field (Manning & Hudgins, 2010), etc. Anomaly is also termed as exceptions, deviations, novelties and noise (Hodge & Austin, 2004). Figure 1 shows the objects which are deviated from the cluster groups C_i in a vector space. The similar objects are grouped as the clusters. The outliers are denoted as $o_1, o_2, \dots, o_n$ and the clusters are denoted as C_i where i denotes the cluster index.

Figure 1. Similarity-based grouping of object with outliers

Commonly, the outliers are detected using the following steps: Initially, a model of the normal activities or behaviour is constructed and then the outliers are detected using the model. The model can be a statistical summary or pattern of the overall samples or objects. The model can be constructed by learning the behaviors or actions of the samples or objects. Using the built model, the outliers are detected by observing the activities or characteristics that significantly differ from the normal model. A model-based outlier detection system with statistical preprocessing is proposed. The statistical measure namely inter-quartile range (IQR) is used for preprocessing and an unsupervised learner is used to build the model in order to detect the outliers for various domains.

Related works

Outliers can be detected in various fashions such as graphical, statistical, unsupervised, supervised, and semi-supervised methods. In graphical method, one or multi-dimensional plots such as spin plot, scatter plot, and box-plot are used to detect and identify the outliers. The major drawback of this approach is the

increased time consumption for detection of the outliers. Also, the plots are more specific to the detection problem, and they follow many constraints.

The statistical method generates a parametric model that describes the data distribution and then, the statistical measure is used to find the parameters of the data distribution such as mean, variance, and median to find the ranges of the outlier confidence level. Then, based on the distribution parameter or outlier ranges, the outliers are detected and separated from the normal distribution of the data model. Zhang et al. (2012) used the statistical method to detect the outliers for the wireless sensor networks (WSN) in order to preprocess the data which is acquired from the sensor nodes for improving the accuracy of the acquired data and to reduce the transmission overhead. In this approach, the statistical measures such as temporal correlation, spatial correlation and spatial-temporal correlation are used to detect the outliers.

Rousseeuw and Hubert (2011) developed an outlier detection scheme using robust location and scatter estimators for outlier detection in multivariate data. The location refers to the coordinate-wise mean and the scatter refers to the covariance matrix. Statistical measure is computed in three phases namely c-step data iteration, data partitioning, and data nesting (Rousseeuw & Hubert, 2011). Hido, Tsuboi, Kashima, Sugiyama, and Kanamori (2011) proposed a statistical-based outlier detection method using the direct density ratio estimation. The major drawback of this statistical method is that most of the statistical tests cannot be applied for the multi-attribute problems. Also, they require the prior knowledge of probability distribution of the data and it is difficult to estimate the real distribution of high dimensional data (Hido et al., 2011).

In the unsupervised method, the data are structured as attributes (column-wise) and objects (row-wise). Objects refer to samples or populations that contain the attributes. Commonly, the clustering approach is used with the similarity criteria such as distance-based, density-based, nearest-neighbor-based, etc. Casas, Mazel, and Owezarski (2011) proposed a network anomaly detection system using the unsupervised clustering approach by sub-space and density-based metrics to prevent the network attacks. The network attacks can be denial-of-service, illegal access, probing, and spreading worms and viruses (Casas et al., 2011). In some cases, multiple-clustering techniques are combined for outlier detection using evidence accumulation (Fred & Jain, 2005). The researchers observed that distance-based method is better in detecting the outliers since it does not require any probability distribution and also computing the probabilistic distribution to the high-dimensional data is difficult.

Sugiyama and Borgwardt (2013) developed an unsupervised outlier detection method using sampling-based in the literature (Sugiyama) and reported that the sampling method outperforms the other method that uses the searching technique using k -nearest neighbor principle. Koupaie, Ibrahim, and Hosseinkhani (2014) suggested unsupervised outlier detection to detect the stream data. The multi-objective genetic algorithm (MOGA) is used to search the outliers from an object space and the k -means clustering is used to develop the model in order to detect the outliers. Casas, Mazel, and Owezarski (2012) presented a knowledge-independent method based on multi-clustering approach, which is adopted to detect the network attack. The anti- k nearest neighbor approach is used to detect the credit card fraud in the literature (Ganji & Mannem, 2012). The prime advantage of the unsupervised outlier detection is that it does not require the labelled data since the labelled data are costlier than unlabelled data and it requires special mechanism to label the data. Therefore, this approach is simple and cost-effective than the supervised approach.

In supervised method, the classifiers are used to detect the outliers by learning the labelled data. The following works were carried out by the researchers in supervised method. A RIPPER classification algorithm is used to describe the logical rules to detect the outliers (Salvador & Chan, 2005) from the temporal time series data. Cabrera, Lewis and Mehra (2001) presented an outlier detection algorithm to detect the network anomaly and Masud, Gao, Khan, Han and Thuraisingham (2011) built an ensemble model with voting approach to detect the class label in order to solve the data stream classification problems. Mourão-Miranda et al. (2011) used one-class support vector machine classifier to detect the outlier patient from a group of patients based on the disease in medical environment. In image processing, the support vector machine classifier is used to detect the changes in very high geometrical resolution (VHR) images (Volpi, Tuia, Bovolo, Kanevski & Bruzzone, 2013). The supervised approach requires the labelled data for detection. Hence, it is expensive than the unsupervised method.

In the semi-supervised method, the labeled and unlabeled data are used to detect the outliers. The semi-supervised approaches are followed by the researchers as reported in the literature (Xue, Shang & Feng, 2010). The authors presented a fuzzy rough c-means clustering to detect the outliers. Noto, Brodley, and Slonim (2012) proposed an outlier detection system. In this system, the normal instances are used to build the ensemble feature to detect the anomaly from the received instances. Daneshpazouh and Sami (2014) used the entropy measure to detect the outliers. Initially, the steadfast negative samples are taken from unlabeled and positive data, and then the outliers are detected based on the

entropy score to remove the outliers. Also, Zamoner and Zhao (2013) presented a score base outlier detection using stochastic network method. A semi-supervised cluster was also proposed in the literature to detect the outliers from the digital mammograms (Thangavel & Mohideen, 2010).

Proposed Model-Based Outlier Detection System with Statistical Preprocessing

Various outlier detection approaches have been developed. The outlier detection requires the domain activity database that is also known as training dataset. This data are classified into two types namely labelled data and unlabelled data. Preparing the labelled data is costlier than the unlabelled data because the special mechanisms are required to prepare the labelled data. Unfortunately, the supervised method fails to process the unlabelled data. In order to overcome this problem, the proposed outlier detection system is developed in an unsupervised fashion. In the unsupervised fashion, the labelled data is not required. Thereby, the initial cost is reduced.

The flowchart representation of the proposed system is depicted in Figure 2. This system consists of three major phases namely removal of outliers, building up of outlier detection model, and detection phase. In removal of outliers phase, the outliers are removed from the training dataset using IQR. In the second phase, the outlier detection model (ODM) is built using the k -means clustering algorithm. In the detection phase, the ODM-based detector identifies the outlier objects from the unknown object space.

The IQR measure is used to preprocess and identify the outliers from the training dataset. The IQR finds the outliers from the dataset by identifying the data which is over ranging from the dataset. The IQR is evaluated as $IQR = Q_3 - Q_1$ where Q_3 and Q_1 are the upper and lower quartiles, respectively. Let us consider a dataset with the numerical data that are arranged in ascending order as 4, 7, 9, 11, 12, and 20. Then, the midpoint is identified to divide the dataset into two halves such as 4, 7, 9 | 11, 12, 20. Then, the median of the lower half and upper half of the data are found. Thus, $Q_1 = 7$ is the median of the lower half data portion and $Q_3 = 12$ is the median of the upper half data portion. Then, the IQR is calculated as $Q_3 - Q_1 = 12 - 7 = 5$. Outlier value $OV = 1.5 \times 5 = 7.5$ where the 1.5 is the outlier factor, lower extreme $LE = Q_1 - OV = 7 - 7.5 = -0.5$ and upper extreme $UE = Q_3 + OV = 12 + 7.5 = 19.5$. Therefore, the normal objects lie in between the upper and lower extreme and if any object lies out of the two (LE and UE) extremes that is treated as outlier.

Figure 2. Flowchart representation of the proposed outlier detection system

The unsupervised k -means (Sammut & Webb, 2011) algorithm is used to build the ODM using the training dataset which is preprocessed using IQR. The ODM is used to categorize the unknown object to detect whether the given object is an outlier or a normal object. In order to build ODM, the number of clusters k is defined in such a way that the k -means algorithm defines a centroid for each cluster. Then, the data points of the training dataset are grouped based on the centroid of each cluster using similarity measure Euclidian distance. The cluster center point is refined in an iterative fashion to form the perfect k groups of data points using the Equation 1.

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(j)=k} \|x_i - x_j\|^2 = \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - m_k\|^2 \quad (1)$$

where $x_1, \dots, x_N$ are the data points of the training dataset. Each observed data point is assigned to any one of the k clusters. $C(i)$ indicates the cluster index of the i^{th} vector. m_k denotes the mean vector of the k^{th} cluster. N_k denotes the number of observations of the k^{th} clusters.

Experimental Setup and Discussion on Result

In order to evaluate the performance of the proposed system, the real-world datasets such as e-mail-based dataset Spambase, communication and control system-based dataset Ionosphere, panel dataset Unbalance, banking dataset Credit-g, and service-based dataset CPU-with-vendor are collected from the UCI repository and other databases (Lichman, 2013).

Table 1 shows the details of datasets used for the experiments. In order to conduct the experiment, initially IQR algorithm is applied on the datasets with the extreme value factor and outlier factors 6.0 and 3.0, respectively. Then, the outliers are detected using IQR algorithm as shown in Table 2. Then, the detected outlier instances are removed from the datasets. In order to build the outlier detection model, the datasets without outliers are fed into the k -means algorithm with the Euclidean distance measure and maximum number of iterations of 500. Then, the outlier detection model is built using k -means algorithm. The unknown objects can be detected using the built model. Then, the test results are obtained to evaluate the performance of the proposed system by observing the sum of squared error (SSE) and time taken to build the model (TBM) for the each number of clusters k from 1 to 10 as tabulated in Table 3 and Table 4.

Table 1. Dataset details

S.No.	Dataset name	No. of instances	No. of attributes
1	Spambase	4601	57
2	Ionosphere	351	34
3	Unbalanced	856	32
4	Credit-g	1000	20
5	CPU-with-vendor	209	9

Figure 3 shows the comparison of SSE without and with IQR on various datasets and it is evident that the proposed approach drastically reduces SSE compared to the outlier detection without IQR. Figure 4 shows the comparison on TBM in second without and with IQR on various datasets and it depicts that the proposed approach drastically reduces the time to build the outlier detection model for various datasets compared to the same without IQR.

MODEL-BASED ODS WITH STATISTICAL PREPROCESSING

Figure 3. Comparison on SSE without and with IQR on various datasets

Figure 4. Comparison on TBM in second without and with IQR on various datasets

Table 2. Identified normal and outlier instances

S.No.	Dataset name	No. of normal instances	No. of outlier instances
1	Spambase	2997	1604
2	Ionosphere	286	65
3	Unbalanced	717	139
4	Credit-g	822	178
5	CPU-with-vendor	171	38

Table 3. Comparison on the SSE in second without and with IQR on various dataset

Dataset	Spambase		Ionosphere		Unbalanced		Credit-g		CPU-with-vendor	
No. of cluster	SSE without IQR	SSE with IQR	SSE without IQR	SSE with IQR	SSE without IQR	SSE with IQR	SSE without IQR	SSE with IQR	SSE without IQR	SSE with IQR
2	805.85	95.77	726.10	381.19	816.98	745.86	5665.99	4514.86	182.01	172.92
3	779.89	89.72	698.25	300.54	704.30	634.68	5429.21	4351.55	169.85	162.98
4	737.68	80.83	585.51	254.90	651.60	532.59	5226.21	4067.71	154.92	153.98
5	718.09	66.25	537.73	234.59	564.56	499.952	5000.33	3981.12	156.57	145.80
6	682.87	58.53	518.41	213.18	497.39	464.60	4755.00	3882.03	143.78	138.12
7	669.83	55.19	512.51	198.76	467.14	423.99	4707.86	3762.71	138.60	128.38
8	597.81	53.91	492.88	181.63	452.88	412.22	4715.21	3705.64	124.80	121.90
9	648.86	51.25	466.82	175.07	424.46	402.31	4583.54	3645.53	119.09	118.00
10	576.90	49.87	452.85	170.46	410.49	376.22	4501.12	3611.47	111.84	109.55

Table 4. Comparison on the TBM in second without and with IQR on various datasets

Dataset	Spambase		Ionosphere		Unbalanced		Credit-g		CPU-with-vendor	
No.of cluster	TBM without IQR	TBM with IQR	TBM without IQR	TBM with IQR	TBM without IQR	TBM with IQR	TBM without IQR	TBM with IQR	TBM without IQR	TBM with IQR
2	0.49	0.14	0.24	0.2	0.38	0.18	0.18	0.17	0.07	0.06
3	5.39	0.18	0.17	0.15	0.44	0.2	0.32	0.32	0.04	0.04
4	5.43	0.23	0.22	0.2	0.76	0.35	0.21	0.13	0.07	0.04
5	6.34	0.24	0.2	0.18	0.72	0.64	0.42	0.19	0.05	0.04
6	6.19	0.25	0.33	0.21	0.68	0.51	0.4	0.19	0.04	0.03
7	6.64	0.26	0.3	0.18	0.56	0.28	0.39	0.36	0.06	0.03
8	7.82	0.26	0.21	0.21	0.49	0.4	0.34	0.26	0.1	0.05
9	7.99	0.26	0.2	0.17	0.94	0.62	0.48	0.32	0.1	0.06
10	8.54	0.27	0.26	0.25	0.56	0.55	0.39	0.33	0.08	0.07

Conclusion

This paper presented a model-based outlier detection system using statistical preprocessing method. The performance of this proposed system is tested on the various real world well-known publically available datasets with the performance evaluation metrics namely sum of squared error (SSE) and time to build model (TBM). From the experimental results, it is observed that the proposed system performs better in terms of SSE and TBM for detecting the outliers regardless of the application domain.

References

- Akhiomen, J. (2013). Data Mining Application for Cyber Credit-Card Fraud Detection System. In P. Perner (Ed.), *Advances in Data Mining. Applications and Theoretical Aspects: 13th Industrial Conference, ICDM 2013, New York, NY, USA, July 16-21, 2013. Proceedings* (pp. 218-228). Berlin, Heidelberg: Springer Berlin Heidelberg
- Cabrera, J. B. D., Lewis, L., & Mehra, R. K. (2001). Detection and classification of intrusions and faults using sequences of system calls. *ACM SIGMOD Record*, 30(4), 25-34. doi:10.1145/604264.604269
- Casas, P., Mazel, J., & Owezarski, P. (2011). UNADA: Unsupervised Network Anomaly Detection Using Sub-space Outliers Ranking. In J. Domingo-Pascual, P. Manzonni, S. Palazzo, A. Pont, & C. Scoglio (Eds.), *NETWORKING*

2011 (pp. 40–51). Springer Berlin Heidelberg. Retrieved September 4, 2014, from http://link.springer.com/chapter/10.1007/978-3-642-20757-0_4

Casas, P., Mazel, J., & Owezarski, P. (2012). Knowledge-independent traffic monitoring: Unsupervised detection of network attacks. *IEEE Network*, 26(1), 13-21. doi:10.1109/MNET.2012.6135851

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys*, 41(3), 1-58. doi:10.1145/1541880.1541882

Daneshpazhouh, A., & Sami, A. (2014). Entropy-based outlier detection using semi-supervised approach with few positive examples. *Pattern Recognition Letters*, 49, 77-84. doi:10.1016/j.patrec.2014.06.012

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6), 835-850. doi:10.1109/TPAMI.2005.113

Fujimaki, R., Yairi, T., & Machida, K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05* (pp. 401-410). New York, NY: ACM. doi:10.1145/1081870.1081917

Ganji, V. R., & Mannem, S. N. P. (2012). Credit card fraud detection using anti-k nearest neighbor algorithm. *International Journal on Computer Science and Engineering*, 4(6), 1035-1039.

Hido, S., Tsuboi, Y., Kashima, H., Sugiyama, M., & Kanamori, T. (2011). Statistical outlier detection using direct density ratio estimation. *Knowledge and Information Systems*, 26(2), 309-336. doi:10.1007/s10115-010-0283-2

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2), 85-126. doi:10.1007/s10462-004-4304-y

Koupaie, M. H., Ibrahim, S., & Hosseinkhani, J. (2014). Outlier detection in stream data by clustering method. *International Journal of Advanced Computer Science and Information Technology (IJACSIT)*, 2(3), 25-34. Retrieved September 4, 2014, from <http://papers.ssrn.com/abstract=2376191>

Lichman, M. (2013). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. Retrieved from: <http://archive.ics.uci.edu/ml>

Manning, M., & Hudgins, L. (2010). Array-based technology and recommendations for utilization in medical genetics practice for detection of

chromosomal abnormalities. *Genetics in Medicine*, 12(11), 742-745.

doi:[10.1097/GIM.0b013e3181f8baad](https://doi.org/10.1097/GIM.0b013e3181f8baad)

Masud, M., Gao, J., Khan, L., Han, J., & Thuraisingham, B. M. (2011). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, 23(6), 859-874. doi:[10.1109/TKDE.2010.61](https://doi.org/10.1109/TKDE.2010.61)

Mourão-Miranda, J., Hardoon, D. R., Hahn, T., Marquand, A. F., Williams, S. C. R., Shawe-Taylor, J., & Brammer, M. (2011). Patient classification as an outlier detection problem: An application of the One-Class Support Vector Machine. *NeuroImage*, 58(3), 793-804. doi:[10.1016/j.neuroimage.2011.06.042](https://doi.org/10.1016/j.neuroimage.2011.06.042)

Noto, K., Brodley, C., & Slonim, D. (2012). FRaC: A feature-modeling approach for semi-supervised and unsupervised anomaly detection. *Data Mining and Knowledge Discovery*, 25(1), 109-133. doi:[10.1007/s10618-011-0234-x](https://doi.org/10.1007/s10618-011-0234-x)

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 73-79. doi:[10.1002/widm.2](https://doi.org/10.1002/widm.2)

Salvador, S., & Chan, P. (2005). Learning states and rules for detecting anomalies in time series. *Applied Intelligence*, 23(3), 241-255. doi:[10.1007/s10489-005-4610-3](https://doi.org/10.1007/s10489-005-4610-3)

Sammur, C., & Webb, G. I. (Eds.). (2011). *Encyclopedia of Machine Learning*. Springer Science & Business Media.

Singhal, S., Gankotiya, A. K., Agarwal, S., & Verma, T. (2012). An investigation of wireless sensor network: A distributed approach in smart environment. *Advanced Computing & Communication Technologies (ACCT), 2012 Second International Conference on*, Rohtak, Haryana, 2012, pp. 522-529. doi:[10.1109/ACCT.2012.22](https://doi.org/10.1109/ACCT.2012.22)

Sugiyama, M., & Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems* 26 (pp. 467-475). Curran Associates, Inc.

Tartakovsky, A. G., Polunchenko, A. S., & Sokolov, G. (2013). Efficient computer network anomaly detection by changepoint detection methods. *IEEE Journal of Selected Topics in Signal Processing*, 7(1), 4-11. doi:[10.1109/JSTSP.2012.2233713](https://doi.org/10.1109/JSTSP.2012.2233713)

Thangavel, K., & Mohideen, A. K. (2010). Semi-supervised k-means clustering for outlier detection in mammogram classification. *Trendz in*

Information Sciences Computing (TISC), 2010 (pp. 68-72). Presented at the Trendz in Information Sciences Computing (TISC), 2010.
doi:[10.1109/TISC.2010.5714611](https://doi.org/10.1109/TISC.2010.5714611)

Volpi, M., Tuia, D., Bovolo, F., Kanevski, M., & Bruzzone, L. (2013). Supervised change detection in VHR images using contextual information and support vector machines. *International Journal of Applied Earth Observation and Geoinformation*, 20, 77-85. doi:[10.1016/j.jag.2011.10.013](https://doi.org/10.1016/j.jag.2011.10.013)

Xue, Z., Shang, Y., & Feng, A. (2010). Semi-supervised outlier detection based on fuzzy rough C-means clustering. *Mathematics and Computers in Simulation*, 80(9), 1911-1921. doi:[10.1016/j.matcom.2010.02.007](https://doi.org/10.1016/j.matcom.2010.02.007)

Zamoner, F. W., & Zhao, L. (2013). A network-based semi-supervised outlier detection technique using particle competition and cooperation. *Intelligent Systems (BRACIS), 2013 Brazilian Conference on* (pp. 225-230).
doi:[10.1109/BRACIS.2013.45](https://doi.org/10.1109/BRACIS.2013.45)

Zhang, Y., Hamm, N. A. S., Meratnia, N., Stein, A., van de Voort, M., & Havinga, P. J. M. (2012). Statistics-based outlier detection for wireless sensor networks. *International Journal of Geographical Information Science*, 26(8), 1373-1392. doi:[10.1080/13658816.2012.654493](https://doi.org/10.1080/13658816.2012.654493)

Analysis and Modeling of Statistical Properties of FMDFB Subband Coefficients

E. Jebamalar Leavline
Anna University
Tiruchirappalli, India

Shunmugam Sutha
University College of Engineering
Panruti, India

Fast Multiscale Directional Filter Bank (FMDFB) is an image representation scheme used in several image processing applications. The statistical nature of the FMDFB subbands is analyzed, and a mathematical model of FMDFB coefficients is proposed. Experimental results are justified by goodness-of-fit tests.

Keywords: Statistical analysis, FMDFB, Subbands, Mathematical modeling, Gaussian distribution

---

Introduction

Statistical analysis plays a vital role in image processing and analysis. In high level image processing such as feature extraction, image analysis, segmentation and object recognition, extraction of statistical features is an important step and these statistical features are used along with the shape features to train the classifier. In other low level image processing methods such as image enhancement and restoration, the statistical analysis is useful to determine the parameters required to perform enhancement and restoration (Hyvärinen, Hurri, & Hoyer, 2009).

The use of multiscale transforms in digital image processing is widely addressed in image processing literature. Fast multiscale directional filter bank is an image representation scheme that is successfully used in feature extraction and image enhancement (Cheng, Law, & Siu, 2007a), (Cheng, Law, & Siu, 2007b). In our previous work, an image denoising scheme with FMDFB is proposed that works well for additive Gaussian and multiplicative speckle noise removal compared to other conventional multiscale denoising methods such as wavelet

---

Dr. E. Jebamalar Leavline is an Assistant Professor in the Department of Electronics and Communication Engineering. Email her at: jebi.lee@gmail.com. Dr. Sutha is an Assistant Professor in the Department of Electrical and Electronics Engineering. Email her at: suthapadmanabhan@gmail.com.

and contourlet based denoising (Leavline, Sutha, & Singh, 2014a). Yet, the statistical nature of the FMDFB subbands has not been analyzed.

Fast Multiscale Direction Filter Banks (FMDFB)

The major drawback of wavelet transform is that it is not being capable of capturing directional information effectively (Po & Do, 2006). This drawback is overcome by employing multiscale and directional representations that capture the geometrical structures in images such as smooth contours (Leavline, Sutha, & Singh, 2014b), (Do & Vetterli, 2005). In spite of high computation complexity and overcompleteness, the multiscale transforms are preferred in various applications because of their ability to represent fine details of the natural images. Pyramidal directional filter banks (PDFB) is one such multiscale image representation scheme, also termed as contourlet. A modification to the PDFB was proposed, namely the multiscale directional filter bank (MDFB) that is redundant in nature and a number of possible structures are available based on the choice of lowpass filter and the number of directional decomposition. According to Cheng et al (2007a), MDFB introduces an additional decomposition in the high-frequency band and thereby improves the radial frequency resolution at a cost of one set of extra scale and directional decompositions on the full image size. This results in increased number of computations. Also, MDFB has a higher redundancy than PDFB. The over completeness and increased frequency resolution of MDFB was found useful in applications like texture characterization and retrieval (Cheng et al., 2007a).

Cheng et al. (2007b) proposed a fast and reduced redundancy structure for this MDFB (FMDFB). This structure has the same redundancy as PDFB and 33% reduction in computational complexity compared to MDFB. Also, FMDFB exhibits perfect reconstruction irrespective of the choice of low pass filters. The total number of FMDFB directional subband coefficients is the same as the size of the original image because of the critically sampled DFB, and hence no extra computations are introduced by the scale decomposition. The computational complexity of FMDFB (Cheng et al., 2007b) is approximated as

$$\left(\frac{19}{3} + \frac{8}{3}L_d\right)MN \quad (1)$$

FMDFB outperforms MDFB in texture retrieval. In our previous work, we have introduced a multiscale denoising approach using FMDFB for Gaussian and speckle noise removal (Leavline et al., 2014a; Leavline & Sutha, 2011).

Statistical Analysis of FMDFB Subbands

Need for Statistical Analysis:

Statistical analysis plays a vital role in all areas of image processing. Although the usefulness of FMDFB has been demonstrated in image processing, it has not been extensively studied. Hence, analysis of statistical nature of the FMDFB subbands and a generalized mathematical model of FMDFB subbands is essential to aid the use of FMDFB in various applications.

Statistical measures

Several statistical measures used to analyze the statistical nature of the FMDFB subbands are presented.

Mean: It is the most common statistical measure as in Equation (2). It gives the average value of intensity level of the image.

$$\mu_x = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N x(i, j) \quad (2)$$

where μ_x is the mean value of the image $x(i, j)$ of size $M \times N$.

Median: It is a measure of much importance as it has a breakdown point of 50%. It represents the mid intensity value present in an image.

Mode: Mode is the most frequently occurring intensity value in $x(i, j)$ of size $M \times N$. However, the mode is not suitable for finding peaks in distributions having multiple modes.

Standard deviation: It is the measure of how the image intensity is spread from its average intensity μ_x . It is calculated as the square root of variance and given as

$$\sigma_x = \sqrt{\frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [x(i, j) - \mu_x]^2} \quad (3)$$

Variance: It is the average of squared difference between $x(i, j)$ and μ_x calculated as square of standard deviation.

$$Var_x = \frac{1}{MN} \sum_{i=1}^M \sum_{j=1}^N [x(i, j) - \mu_x]^2 \quad (4)$$

Geometric mean: The geometric mean is less than or equal to the arithmetic mean unless all the intensity levels equal. It expresses the central tendency of the intensity values using the product of their values. The geometric mean is generally defined as the n^{th} root of the product of ‘ n ’ numbers.

$$GM_x = \left[\prod_{i,j=1}^{M,N} x(i, j) \right]^{1/MN} \quad (5)$$

Harmonic mean: Harmonic mean is calculated as the reciprocal of the arithmetic mean of the reciprocals. It is less than or equal to arithmetic mean. It is also called as sub-contrary mean.

$$HM_x = \frac{MN}{\sum_{i,j=1}^{M,N} \frac{1}{x(i, j)}} \quad (6)$$

Trimmed-mean: The trimmed-mean is a robust estimate of the location of a pixel $x(i, j)$. It calculates the mean by excluding ‘ K ’ smallest and largest intensity values. It is a representative estimate of the central intensity values of the image in the presence of outlier intensities. If $t(k, l)$ represents the trimmed intensity values with $k = 1, 2, \dots, K$ and $l = 1, 2, \dots, L$ the trimmed-mean is calculated as

$$TM_x = \frac{\sum_{k,l=1}^{K,L} t(k, l)}{KL} \quad (7)$$

Range: Range is the difference between minimum and maximum intensity level present in the image.

$$Range_x = \max(x(i, j)) - \min(x(i, j)) \quad (8)$$

Inter-quartile range: It is also known as midspread and is calculated as the difference between the 75th and the 25th percentiles of the intensity values of the image.

Mean absolute deviation: It is also termed as mean deviation or average absolute deviation. It is the mean of absolute deviation of a particular intensity level from the mean intensity level. This measure is computationally efficient than standard deviation.

$$MAD = \frac{1}{MN} \left[\sum |x(i, j) - \mu_x| \right] \quad (9)$$

Gray Level Co-occurrence Matrix (GLCM): It is calculated as how often a pixel with gray-level (grayscale intensity) value 'g' occurs horizontally adjacent to a pixel with the value 'h'. Another name for a gray-level co-occurrence matrix is a gray-level spatial dependence matrix. The measures such as contrast, energy, homogeneity, entropy, and correlation are calculated from normalized GLCM.

Contrast: It is intensity contrast between a pixel and its neighbor over the whole image.

$$C_x = \sum_{g,h} |g - h|^2 GLCM(g, h) \quad (10)$$

Energy: It is the sum of squared elements in the GLCM.

$$E_x = \sum_{g,h} [GLCM(g, h)]^2 \quad (11)$$

Homogeneity: It is the measure of the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

$$H_x = \sum_{g,h} \frac{GLCM(g,h)}{1+|g-h|} \quad (12)$$

Entropy: Entropy is a statistical measure of randomness that can be used to characterize the texture of the input image.

$$Entropy_x = -\sum p(x) \log_2 p(x) \quad (13)$$

Correlation: Correlation between two images $x(i, j)$ and $y(i, j)$ is calculated as

$$r = \frac{\sum_i \sum_j (x(i, j) - \mu_x)(y(i, j) - \mu_y)}{\sqrt{\left(\sum_i \sum_j x(i, j) - \mu_x\right)^2 \left(\sum_i \sum_j y(i, j) - \mu_y\right)^2}} \quad (14)$$

Moments: The central moment of order ‘ n ’ is defined as

$$m_n = E(x - \mu)^n \quad (15)$$

The central first moment is zero, and the second central moment is the variance. The third and fourth central moments are used to find the skewness and kurtosis. The normalized n^{th} central moment is calculated as the n^{th} central moment divided by σ^n .

Mathematical Modeling of FMDFB coefficients

A statistical model for FMDFB transform coefficients will now be developed. A computationally efficient and accurate model of FMDFB coefficients is necessary to assist straightforward parameter estimation needed for various image processing applications (Sutha, Leavline, & Gnana Singh, 2013). First, to find a suitable statistical model (Kwitt, 2010), the frequency distribution of FMDFB subbands need to be analyzed. For this purpose, we use the classical histogram. Also, the histogram is plotted along with the Gaussian fit function that is shown red in color in Figure 1 to Figure 4.

The goodness-of-fit (GoF) of the FMDFB subband coefficients with the Gaussian distribution is evaluated with the help of Quantile–Quantile plot (Q-Q

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

plot) as a graphical tool. The Q-Q plots of FMDFB subbands of a clean image shows that, except the tail regions at both the ends, the FMDFB subbands coincide with the Gaussian (normal) distribution shown red in color in Figure 5 and 6. However, it is evident from the Q-Q plots of the FMDFB subbands of noisy image that they overlap with the Gaussian (normal) distribution at most of the points as in Figure 7 and 8.

Figure 1. Histogram of level 1 FMDFB subbands of clean Lena image

Figure 2. Histogram of level 1 FMDFB subbands of Lena image with AWGN with variance 0.001

Figure 3. Histogram of level 2 FMDFB subbands of clean Lena image

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Figure 4. Histogram of level 2 FMDFB subbands of Lena image with AWGN with variance 0.001

Figure 5. Quantile – Quantile plot of level 1 FMDFB subband coefficients of clean Lena image

Figure 6. Quantile – Quantile plot of sample level 2 FMDFB subband coefficients of clean Lena image

Figure 7. Quantile – Quantile plot of level 1 FMDFB subband coefficients of Lena image with AWGN with variance 0.001

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Figure 8. Quantile–Quantile plot of sample level 2 FMDFB subband coefficients of Lena image with AWGN with variance 0.001

To add strength to this argument, the GoF is tested using the Chi-Squared test. The MATLAB (“MATLAB”) statistical toolbox function $[h, p, stat] = chi2gof(x)$ is used that performs a chi-square goodness-of-fit test of the default null hypothesis that the data in vector ‘ x ’ are a random sample from a normal distribution with mean and variance estimated from ‘ x ’, against the alternative that the data are not normally distributed with the estimated mean and variance. The result ‘ h ’ is 1 if the null hypothesis can be rejected at the 5% significance level denoted as ‘ α ’. The result ‘ h ’ is 0 if the null hypothesis cannot be rejected at the 5% significance level. The ‘ p ’ value is the probability, under assumption of the null hypothesis of observing the given statistic and ‘ $stat$ ’ is the chi-squared statistics. The following procedure is followed to perform the chi-square goodness-of-fit test.

- Step 1. The image is read and decomposed using FMDFB into subbands. In this experiment, the scale (s) and directional (l) decomposition

is fixed as 2. For this specification, there will be four and sixteen subbands at level 1 and level 2 of decomposition respectively. We denote the level 1 subband as $SB_{ij}, i = 1, 2 \dots 2^s, j = 0$ and level 2 subbands as $SB_{ij}, i = 1, 2 \dots 2^s, j = 1, 2, \dots, 2^1$

- Step 2. For a subband SB_{ij} , the parameters of Gaussian (normal) distribution mean (μ) and standard deviation (σ) are estimated using maximum likelihood estimator with 95% confidence interval.
- Step 3. Using the estimated ' μ ' and ' σ ', a reference Gaussian (normal) distribution is generated. From this distribution, ' N ' samples are selected uniformly at random without replacement.
- Step 4. Then the chi-square goodness-of-fit is calculated. If the ' $Chi2Stat$ ' is less than the critical value calculated from the chi square table for the particular degrees of freedom the null hypothesis is accepted and it is rejected otherwise following the test hypothesis given below.

- H_0 . The coefficients of FMDFB subbands are consistent with the Gaussian (normal) distribution.
- H_a . The coefficients of FMDFB subbands are not consistent with the Gaussian (normal) distribution.

The h , p and $stat$ of the FMDFB subbands are indicted in Table 1. ' h ' represents whether the null hypothesis is accepted or rejected. ' p ' represents the probability with which the test hypothesis is accepted. If ' p ' is less than 0.05, then the null hypothesis is rejected. The parameter ' $stat$ ' contains the chi square test value and the degrees of freedom. It is evident from Table 1 that the null hypothesis is rejected only for a few subbands. Further, the above mentioned test is also conducted on various standard gray scale images. The average rate of hypothesis acceptance is 96.66% as shown in Table 2. Hence, the Gaussian probability distribution assumption of FMDFB subbands holds good.

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Table 1. Results of Chi-square test of FMDFB subbands

Subbands	Clean Lena Image				Noisy Lena Image with AWGN 0.001				Noisy Lena Image with Speckle 0.04			
	<i>h</i>	<i>p</i>	<i>Chi2</i> value	DoF	<i>h</i>	<i>p</i>	<i>Chi2</i> value	DoF	<i>h</i>	<i>p</i>	<i>Chi2</i> value	DoF
SB10	0	0.0619	8.966	4	0	0.6666	4.0746	6	0	0.7094	3.7582	6
SB20	0	0.8250	3.595	7	0	0.6189	4.4283	6	1	0.0106	16.663	6
SB30	0	0.0629	13.403	7	0	0.3055	8.3162	7	0	0.4252	7.0357	7
SB40	0	0.9880	0.934	6	0	0.2133	7.0999	5	0	0.3602	5.4796	5
SB11	0	0.8685	3.174	7	0	0.5592	3.9322	5	0	0.7147	2.9049	5
SB12	0	0.4215	6.014	6	0	0.8168	2.9363	6	0	0.4700	5.5950	6
SB13	0	0.4504	5.762	6	1	0.0219	13.1671	5	0	0.8535	3.3245	7
SB14	0	0.6165	4.446	6	0	0.3267	5.7949	5	0	0.1704	7.7525	5
SB21	0	0.2525	9.002	7	1	0.0170	13.7923	5	0	0.5547	4.9158	6
SB22	0	0.0883	12.394	7	0	0.7317	2.7943	5	0	0.9317	2.4384	7
SB23	0	0.4344	5.900	6	0	0.7267	2.0492	4	0	0.7336	3.5775	6
SB24	0	0.8166	3.672	7	1	0.0202	15.0073	6	0	0.2490	9.0512	7
SB31	0	0.7600	3.379	6	0	0.2208	6.9979	5	0	0.6080	4.5101	6
SB32	1	0.0339	13.644	6	0	0.2470	9.0794	7	0	0.5819	4.7068	6
SB33	0	0.3148	7.065	6	0	0.9336	2.4122	7	0	0.3610	6.5839	6
SB34	1	0.0144	17.504	7	0	0.3168	5.8930	5	0	0.4185	7.1002	7
SB41	0	0.4566	5.708	6	0	0.2594	8.9073	7	0	0.2283	8.1362	6
SB42	0	0.4582	5.695	6	0	0.4941	6.3979	7	0	0.5778	5.6780	7
SB43	0	0.7084	3.765	6	0	0.6206	5.3228	7	0	0.1495	8.1237	5
SB44	0	0.0870	11.045	6	0	0.4728	5.5716	6	0	0.8157	2.9453	6

Table 2. Summary of chi-square goodness of fit test for 20 gray scale images

Image	Clean Image		Image with Gaussian Noise (zero mean & 0.001 variance)		Image with Speckle Noise (0.04 variance)	
	Accepted Subbands	Rejected Subbands	Accepted Subbands	Rejected Subbands	Accepted Subbands	Rejected Subbands
Total	385	15	390	10	385	15
Acceptance (%)	96.25		97.5		96.25	

FMDFB Subband Model:

FMDFB is an orthogonal filter bank structure that is critically sampled in nature. It is a multiresolution and directional scheme that is useful for image representation. In order to visualize the nature of the FMDFB subband coefficients, the histogram fit and chi-square goodness-of-fit test are performed. From the discussions presented in the previous section, it is concluded that the FMDFB subbands follow the Gaussian (Normal) distribution with mean ' μ ' and standard deviation ' σ '. Following that, the FMDFB coefficients are mathematically modeled with the following assumptions.

- Assumption 1. The coefficients of any FMDFB subband are identically distributed with the same probability density function.
- Assumption 2. The coefficients of FMDFB subbands of the same level of decomposition are not independent and they are highly correlated.
- Assumption 3. The coefficients of FMDFB subbands of different levels of decomposition are independent and they are not correlated.

With a scale decomposition of ' s ', and level of directional decomposition ' l ', the FMDFB subband coefficients are represented as in Equation (16) where ' γ ' is the FMDFB subband.

$$\psi^{(l,s)} = \begin{cases} \{\gamma_i\}, i=1,2,\dots,2^l, & \text{for } s=1 \\ \{\gamma_{ij}\}, i=1,2,\dots,2^l, j=1,2,\dots,2^s & \text{for } s \neq 1 \end{cases} \quad (16)$$

The standard Gaussian (normal) distribution of a random function ‘g’ follows the probability density function given by Equation (17) with mean $\mu = 0$ and variance $\sigma = 1$.

$$f(g, \mu, \sigma) = \frac{1}{\sqrt{2\pi}} e^{-\frac{g^2}{2}} \quad (17)$$

Then the probability density function $\zeta(.)$ of the FMDFB coefficients is modeled as in Equation (18). Here, γ_{ij} are the coefficients of FMDFB subband, μ_γ and σ_γ are the mean and standard deviation of the FMDFB subband under consideration.

$$\zeta(\gamma_{ij}, \mu_\gamma, \sigma_\gamma) = \frac{1}{\sigma_\gamma \sqrt{2\pi}} e^{-\frac{(\gamma_{ij} - \mu_\gamma)^2}{2\sigma_\gamma^2}} \quad (18)$$

Results and Discussion

Experimental Setup:

The analysis of statistical characteristics of FMDFB subbands is carried out in MATLAB environment on standard test images. The statistical measures are calculated on images with the help of MATLAB Statistics Toolbox and Image Processing Toolbox.

Results and Discussion:

The statistical measures are calculated on clean image and noisy image. The comparison of various statistical measures is shown in Table 3 for various noise densities. Also, the noisy image is decomposed using FMDFB with scale and directional decomposition levels set to 2. This will result in 16 subbands. Further, the statistical measures of all the sixteen FMDFB directional subband coefficients are calculated and tabulated in Table 4. Correlation between pairs of subbands is studied at level 1 and level 2 of FMDFB decomposition as shown in Table 5 and Table 6. An analysis of first, second and higher order moments is carried out and the moments for level 1 and level 2 FMDFB subbands are shown in Figure 9. The histograms of level 1 and level 2 FMDFB subbands of clean and noisy image with speckle noise are shown in Figure 10 and Figure 11.

The following facts are inferred from the experimental results.

- (i) The mean, standard deviation and mean absolute deviation (MAD) of the image vary linearly with the level of noise present.
- (ii) The geometric mean and harmonic mean are non zero for clean image and they become zero for noisy image. However, at some higher noise levels, these two statistical parameters become non zero.
- (iii) Energy and homogeneity of the intensity values are almost constant and equal to 1. But entropy is non zero and is not constant with respect to the level of noise present.
- (iv) After performing FMDFB decomposition, the subband SB11 exhibits maximum magnitudes for all statistical measures compared at a particular noise level. On the other hand the subband SB33 exhibits maximum magnitudes for all statistical measures compared.

Figure 9. First, Second and higher order moments of FMDFB Subbands

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Figure 10. Histogram of level 1 FMDFB subbands of Lena image with speckle noise with variance 0.04

Figure 11. Histogram of level 2 FMDFB subbands of Lena image with speckle noise with variance 0.04

- (v) However, the energy and homogeneity are constant and equal to 1 for all subbands irrespective of the level of decomposition, and the entropy is maintained at zero. Energy is equal to 1 for clean images and the homogeneity of magnitude 1 represents that all the subbands are smooth in nature. Also, lower entropy values depict that the subbands are having smooth texture.
- (vi) Among the level 1 subbands, {SB3, SB4} subband pair exhibits less correlation compared to other pairs of subbands. However, the correlation coefficients of all level 1 subbands are close to unity, showing higher degree of correlation.
- (vii) In level 2, the subbands {SB11, SB33} subband pair exhibits less correlation.
- (viii) From the analysis of lower and higher order central moments, it is observed that, all the subbands have the first order central moment as zero. This shows that all the FMDFB subbands follow a distribution with zero mean. The second moment is the variance of each subband and SB11 has maximum variance meaning that the coefficients of SB11 subband are spread over a wide range.
- (ix) The third normalized central moment is a measure of skewness of the histogram. The coefficients of all FMDFB subbands have negative skewness. This means that the histograms of the subbands are skewed towards the left. However, since the skewness is near zero, the distribution of the subband coefficients can be characterized as near symmetric. The fourth normalized central moment is a measure of shape of the distribution. The kurtosis values of all the subbands are above zero (positive). This ensures that there will be a peak in the distribution which falls at the mean. The average magnitude of excess kurtosis of subbands is near zero. Hence the shape of the distribution can be characterized as Gaussian (Normal). Also, the average proper kurtosis of FMDFB subbands of a noisy image is approximately 3 leading to Gaussian (Normal) distribution. However, the average proper kurtosis of FMDFB subbands of a clean image is less than 3 leading to sub-Gaussian (sub-Normal) distribution or Platykurtic distribution.
- (x) The images corrupted with Gaussian noise and speckle noise follow similar probability distribution as in [Figure 10](#) and [Figure 11](#).

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Table 3. Comparison of various statistical measures of clean and noisy images

Gaussian Noise Variance	Mean	Median	Mode	SD	Variance	GM	HM	TM	IQR	MAD	Range	Contrast	Corre- lation	Energy	Homo- geneity	Entropy
Clean Image	123.61	140	99	12.531	822850	112.52	99.84	124.11	35	9.418	193	0	-	1	1	0
0.001	123.94	133	0	9.667	812257.4	0	0	124.37	27.25	7.221	117	0.530	0.036	0.978	0.991	0.050
0.005	124.85	134	0	9.695	808297.7	0	0	125.29	27.25	7.232	115	0.475	0.029	0.980	0.992	0.046
0.01	126.10	136	0	9.732	819356	0	0	126.60	27	7.256	125	0.436	0.030	0.982	0.992	0.042
0.02	128.68	138.25	255	9.576	794593.6	0	0	129.15	27.5	7.148	112	0.355	0.023	0.985	0.994	0.035
0.03	131.12	140.25	255	9.670	809828.5	0	0	131.69	28	7.200	118	0.272	0.021	0.989	0.995	0.028
0.04	133.76	143	255	9.631	798620.9	0	0	134.32	28	7.244	114	0.223	0.024	0.991	0.996	0.024
0.05	136.24	145.75	255	9.645	804415.7	0	0	136.86	27.5	7.208	119	0.178	0.017	0.993	0.997	0.020
0.06	138.79	148.5	255	9.625	801055.4	0	0	139.37	27.5	7.197	119	0.139	0.009	0.994	0.998	0.016
0.07	141.25	150.25	255	9.611	789655	0	0	141.80	27	7.204	124	0.099	-0.001	0.996	0.998	0.012
0.08	143.79	153	255	9.522	773594.8	0	0	144.49	28.75	7.203	122	0.075	0.009	0.997	0.999	0.009
0.09	146.32	155.75	255	9.511	769365.9	0	0	146.94	29	7.200	126	0.052	0.007	0.998	0.999	0.007
0.1	148.85	158.25	255	9.496	767611.6	0	0	149.60	27.5	7.129	134	0.040	0.000	0.998	0.999	0.005
0.2	173.32	184.75	255	9.009	620831.8	0	0	175.15	28.5	6.958	156	0.002	0.000	1.000	1.000	0.000

EPIPHANY & SHUNMUGAM

Table 4. Comparison of various statistical measures of different FMDFB subbands (level 1 and level 2)

FMDFB Subbands	Mean	Median	Mode	Standard Deviation	GM	HM	TM	IQR	MAD	Range
SB1	247.5	277.3	54.2	24.8	229.1	208.7	248.6	69.4	18.5	323.4
SB2	124.0	138.2	30.7	12.4	114.8	104.5	124.6	36.0	9.3	160.7
SB3	62.0	69.1	12.6	6.0	57.4	52.1	62.2	17.4	4.5	82.0
SB4	123.8	137.6	28.4	12.2	114.4	104.1	124.2	35.0	9.1	163.7
SB11	495.1	556.7	100.8	50.0	458.3	417.4	497.1	137.3	37.0	640.8
SB12	247.5	277.4	61.8	25.0	228.8	207.9	248.6	69.1	18.7	324.6
SB13	123.8	138.7	30.8	12.4	114.6	104.3	124.3	33.9	9.3	160.4
SB14	247.5	276.8	55.5	24.8	229.5	209.5	248.6	70.0	18.3	317.7
SB21	248.0	276.5	61.8	25.0	229.6	209.2	249.1	72.1	18.7	311.3
SB22	124.0	138.1	27.8	12.5	114.6	104.0	124.6	36.4	9.4	163.8
SB23	62.0	69.2	14.4	6.2	57.4	52.2	62.3	18.0	4.7	80.1
SB24	124.0	139.0	32.9	12.4	115.0	105.0	124.6	36.8	9.3	158.6
SB31	124.0	138.2	25.0	12.1	114.7	104.4	124.4	35.4	9.1	170.0
SB32	62.0	69.1	12.4	6.1	57.2	51.9	62.2	18.1	4.6	82.3
SB33	31.0	34.3	6.2	3.0	28.7	26.0	31.1	9.0	2.3	40.2
SB34	62.0	68.7	12.0	6.0	57.4	52.4	62.2	17.6	4.5	79.9
SB41	247.5	275.5	58.1	24.6	228.9	208.2	248.5	71.1	18.2	320.6
SB42	123.8	137.2	26.1	12.3	114.3	103.7	124.3	35.8	9.2	153.7
SB43	61.9	68.5	13.4	6.1	57.2	52.0	62.1	16.9	4.6	78.2
SB44	123.8	137.4	26.5	12.2	114.6	104.5	124.3	34.6	9.0	164.1

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Table 5. Comparison of correlation between FMDFB subbands at level 1 (clean image)

Level 1 Subbands	SB1	SB2	SB3	SB4
SB1	1.00	0.98	0.97	0.98
SB2	0.98	1.00	0.98	0.98
SB3	0.97	0.98	1.00	0.96
SB4	0.98	0.98	0.96	1.00

Table 6. Comparison of correlation between FMDFB subbands at level 2 (clean image)

Level 2 Subbands	SB11	SB12	SB13	SB14	SB21	SB22	SB23	SB24	SB31	SB32	SB33	SB34	SB41	SB42	SB43	SB44
SB11	1.00	0.97	0.94	0.99	0.99	0.93	0.90	0.97	0.97	0.91	0.88	0.95	0.98	0.95	0.92	0.97
SB12	0.97	1.00	0.99	0.98	0.98	0.98	0.96	0.98	0.98	0.97	0.94	0.97	0.96	0.98	0.96	0.96
SB13	0.94	0.99	1.00	0.97	0.97	0.99	0.98	0.98	0.97	0.98	0.96	0.98	0.94	0.98	0.98	0.96
SB14	0.99	0.98	0.97	1.00	0.99	0.95	0.93	0.98	0.98	0.93	0.91	0.97	0.98	0.96	0.95	0.98
SB21	0.99	0.98	0.97	0.99	1.00	0.96	0.94	0.99	0.98	0.94	0.92	0.97	0.98	0.97	0.95	0.97
SB22	0.93	0.98	0.99	0.95	0.96	1.00	0.99	0.98	0.96	0.98	0.96	0.96	0.93	0.98	0.97	0.94
SB23	0.90	0.96	0.98	0.93	0.94	0.99	1.00	0.97	0.94	0.98	0.98	0.96	0.91	0.96	0.98	0.93
SB24	0.97	0.98	0.98	0.99	0.99	0.98	0.97	1.00	0.98	0.96	0.94	0.98	0.97	0.98	0.97	0.98
SB31	0.97	0.98	0.97	0.98	0.98	0.96	0.94	0.98	1.00	0.96	0.94	0.99	0.96	0.96	0.95	0.96
SB32	0.91	0.97	0.98	0.93	0.94	0.98	0.98	0.96	0.96	1.00	0.99	0.98	0.91	0.96	0.96	0.93
SB33	0.88	0.94	0.96	0.91	0.91	0.96	0.98	0.94	0.94	0.99	1.00	0.96	0.89	0.94	0.96	0.91
SB34	0.95	0.97	0.98	0.97	0.97	0.96	0.96	0.98	0.99	0.98	0.96	1.00	0.95	0.96	0.96	0.96
SB41	0.98	0.96	0.94	0.98	0.98	0.93	0.91	0.97	0.96	0.91	0.89	0.95	1.00	0.96	0.94	0.99
SB42	0.945	0.979	0.978	0.960	0.967	0.976	0.965	0.975	0.962	0.959	0.942	0.963	0.964	1.000	0.988	0.977
SB43	0.919	0.963	0.977	0.943	0.945	0.972	0.976	0.965	0.945	0.963	0.960	0.959	0.937	0.988	1.000	0.964
SB44	0.968	0.962	0.960	0.980	0.972	0.941	0.931	0.978	0.963	0.925	0.909	0.961	0.990	0.977	0.965	1.000

Conclusion

Experimental results revealed the FMDFB directional subbands follow Gaussian (normal) probability distribution with mean ' μ ' and standard deviation ' σ '. This statistical model will be very much useful to estimate necessary parameters for image processing applications such as threshold estimation in denoising and segmentation and feature extraction.

References

- Cheng, K.-O., Law, N.-F., & Siu, W.-C. (2007a). Multiscale directional filter bank with applications to structured and random texture retrieval. *Pattern Recognition*, 40(4), 1182-1194. doi:10.1016/j.patcog.2006.07.014
- Cheng, K.-O., Law, N.-F., & Siu, W.-C. (2007b). A novel fast and reduced redundancy structure for multiscale directional filter banks. *IEEE Transactions on Image Processing*, 16(8), 2058-2068. doi:10.1109/TIP.2007.901212
- Do, M. N., & Vetterli, M. (2005). The contourlet transform: An efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12), 2091-2106. doi:10.1109/TIP.2005.859376
- Hyvärinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural image statistics: A probabilistic approach to early computational vision*. (Vol. 39). New York, NY: Springer. Retrieved May 27, 2014
- Kwitt, R. (2010). *Statistical modeling in the wavelet domain and applications* (Doctoral thesis). University of Cambridge, Cambridge, United Kingdom. Retrieved May 27, 2014, from <http://wavelab.at/papers/Kwitt10d.pdf>
- Leavline, E. J., & Sutha, S. (2011, June 3-5). Gaussian noise removal in gray scale images using fast multiscale directional filter banks. *Recent Trends in Information Technology (ICRTIT), 2011 International Conference on* (pp. 884-889). Presented at the 2011 International Conference on Recent Trends in Information Technology (ICRTIT), Chennai, India. doi:10.1109/ICRTIT.2011.5972377
- Leavline, E. J., Sutha, S., & Singh, D. A. A. G. (2014a). Fast multiscale directional filter bank-based speckle mitigation in gallstone ultrasound images. *Journal of the Optical Society of America A*, 31(2), 283-292. doi:10.1364/JOSAA.31.000283

STATISTICAL ANALYSIS OF FMDFB SUBBAND COEFFICIENTS

Leavline, E. J., Sutha, S., & Singh, D. A. A. G. (2014b). On the suitability of multiscale image representation schemes as applied to noise removal. *International Journal of Innovative Computing, Information and Control*, 10(3), 1135-1147.

MATLAB – The Language of Technical Computing – MathWorks India. Retrieved May 27, 2014, from <http://www.mathworks.in/products/matlab/>

Po, D. D.-Y., & Do, M. N. (2006). Directional multiscale modeling of images using the contourlet transform. *IEEE Transactions on Image Processing*, 15(6), 1610-1620. doi:10.1109/TIP.2006.873450

Sutha, S., Leavline, E. J., & Gnana Singh, D. A. A. (2013). IHNS: A pragmatic investigation on identifying highly noisy subband in FMDFB for fixing threshold to deteriorate noise in images. *Information Technology Journal*, 12(7), 1289-1298. doi:10.3923/itj.2013.1289.1298

An Evaluation of Pareto, Lognormal and PPS Distributions: The Size Distribution of Cities in Kerala, India

Christopher A. Vallabados
S.R.M. Medical College and Research Centre
Kattankulathur, India

Subbarayan A. Arumugam
S.R.M. University
Kattankulathur, India

The Pareto-Positive Stable (PPS) distribution is introduced as a new model for describing city size data of a region in a country. The PPS distribution provides a flexible model for fitting the entire range of a set of city size data and the classical Pareto and Zipf distributions are included as a particular case.

Keywords: City-size distribution, Pareto distribution, log normal distribution, Zipf's law, positive stable law

---

Introduction

Systems with measurable entities (which can be defined by their size) are characterized by particular properties of their distribution. There are extensive literature and case studies in this field that include work on population of countries, incomes of people in the same economy, frequency of words in languages etc. Scholars have been addressing the problem, regarding the size distribution of such systems; the first is finding a mathematical description for these distributions. The most popular suggestions are the lognormal distribution and the power law (known also as Zipf's law). Yet, there are other expressions that describe with equal success general observed distributions. The second problem is to develop model, which explains the size distribution. Here also several models (either analytical or computer simulations) were proposed. These models can be divided into two classes: the first includes models with a limited number of parameters, and the second class includes mostly economic models which are more complex and includes numerous parameters.

Dr. Vallabados is a Lecturer in the Department of Statistics. Email him at: christopheramalraj@gmail.com. Dr. Arumugam is Professor and Head of the Department of Computer Applications. Email him at: subbarayan1948@gmail.com.

Pareto Distribution

The linear relation between population of cities and their ranks on a log-log plot is found to be a power law, where the absolute value of this linear function is the exponent of the power law. A power law is also known as a classical Pareto distribution with cumulative distribution function (cdf),

$$F(x) = P_r(X \leq x) = 1 - \left(\frac{x}{\sigma}\right)^{-\alpha}, x \geq \sigma > 0 \text{ and } F(x) = 0 \text{ if } x < \sigma, \quad (1)$$

where $\alpha > 0$ is a shape parameter and σ is a scale parameter, which represents the population of the smallest city in the sample. The α parameter is called the Pareto coefficient. The quantity $\left(\frac{x}{\sigma}\right)^{-\alpha}$ represents the proportion of cities of large size than a given x value.

A Select Review of City Size Distribution Models

Pareto distribution was initially proposed Auerbach (1913) and followed by Zipf (1949) to fit city size data. Rosen and Resnick (1980) did a cross-country investigation of city sizes in 44 countries and found that Pareto exponent was in the interval $\alpha \in [0.81 \text{ to } 1.96]$. They have also tried to explain the variations in the Pareto exponent, and showed that it is sensitive to city definition and city sample size. Based on 135 USA metropolitan areas in 1991, Krugman (1996) calculated the value of α close to one. Using the same data set, Gabaix (1999a, 1999b) derived a statistical explanation of Zipf's law for cities. Brakman, Garretsen, Van Marrewijk, & Van Den Berg (1999) with Netherland data provided Pareto evidence over a wide range of time. Nitsch (2005) used meta analysis and concluded that Pareto distribution as an appropriate one to fit city size data. Zanette and Manrubia (1997) developed an intermittency model to large-scale city size distributions. Davis and Weinstein (2002) found that variation in Japanese regional population density, as well as the distribution of city sizes, obeyed a Pareto distribution, at all points in time. Soo (2005) updated α values for the internal $[0.73, 1.72]$ and tried to explain variations in the Pareto exponent. Moura and Riberio (2006) have showed that Pareto distribution was not valid for smaller cities.

Some probabilistic and economic models have been proposed by many researchers, and the central idea among the above models is that Gibart's law

(proportional growth) can lead to Pareto distribution. Simon (1955) has shown that a proportional growth can explain several different skew distributions, including lognormal, Pareto and Yule. Anderson and Ge (2005) have shown the superiority of the lognormal distribution with respect to Pareto distribution, using size distribution of Chinese cities. Subbarayan (2009) extensively studied the size distribution of cities in Tamilnadu, Indian state for the period 1901-2001. Sarabia and Prieto (2009) have stated that the validity of the Pareto distribution disappears when all the population is fitted, including cities of medium and small size.

The models considered here evolved by Sarabia and Prieto (2009). The descriptive model evolved by them is called PPS distribution for city / town size data. More flexible models emerge from PPS under certain conditions. The classical Pareto and Zipf distributions are included as particular cases. The PPS distribution provides a flexible model for fitting the entire range of a set of city / town size data, when zero and uni-modality are possible (i.e., the probability density function always decreases or it has a local maximum)

The PPS Distribution

Sarabia and Prieto (2009) defined PPS distribution in terms of cdf.

$$\begin{aligned} \text{If } F(x) &= P_r(X \leq x), \text{ then} \\ F(x) &= 1 - \exp\left\{-\lambda \left[\log(x/\sigma)\right]^v\right\}, x \geq \sigma \text{ and} \\ F(x) &= 0 \text{ if } x < \sigma, \text{ where } \lambda, \sigma, v > 0 \end{aligned} \quad (2)$$

A random variable with cdf given by (2) will be denoted by $X \sim \text{PPS}(\lambda, \sigma, v)$. It may be noted that λ and v are shape parameters and σ is a scale parameter.

- Zipf distribution ($\lambda = v = 1$)
- Classical Pareto ($v = 1$)

More flexible models emerge when $v > 1$.

PPS based on Weibull Distribution

PPS distribution can also be obtained from a monotonic transformation of the Weibull distribution.

Let Z be a classical Weibull distribution with cdf

SIZE DISTRIBUTION OF CITIES IN KERALA

$$F_z(z) = 1 - \exp(-z^\nu), z > 0, \text{ where } \nu > 0 \quad (3)$$

then the random variable

$$X = \sigma \exp[l^{\frac{1}{\nu}} Z] \quad (4)$$

where $\sigma, \lambda > 0$ is distributed according to a PPS (λ, σ, ν) distribution with cdf by (2). Using Eq.(4), if X is a PPS distribution with cdf given by Eq.(2), the random variable.

$$Z = l^{1/\nu} \log(x / \sigma)$$

is a Weibull random variable with cdf by (3).

The pdf of PPS is given by

$$f(x) = \frac{dF(x)}{dx} = \frac{\lambda \nu \left[\log\left(\frac{x}{\sigma}\right) \right]^{\nu-1}}{x} \exp \left\{ -\lambda \left[\log\left(\frac{x}{\sigma}\right) \right]^\nu \right\}, x \geq \sigma \quad (5)$$

and $f(x) = 0$ if $x < \sigma$.

If $\nu > 1$ the mode (a local maximum of the pdf) defined by Eq.(5) is at $\sigma \exp(z_0)$, where z_0 is the unique solution of the equation in z ,

$$\lambda \nu Z^\nu + Z - (\nu - 1) = 0$$

Three-parameter Lognormal Distribution

The pdf of the three-parameter lognormal distribution

$$f(x; \mu, \sigma, \gamma) = \frac{1}{(x - \gamma) \sigma \sqrt{2\pi}} \exp \left\{ -\frac{[\ln(x - \gamma) - \mu]^2}{2\sigma^2} \right\} \quad (6)$$

where $x > \gamma \geq 0$, $-\infty < \mu < \infty$, $\sigma > 0$ and γ is the threshold parameter or location parameter that defines the point where the support set of the distribution begins; μ is the scale parameter that stretch or shrink the distribution and σ is the shape parameter that affects the shape of the distribution.

If X is a random variable that has a three parameter log-normal probability distribution, then $Y = \ln(X - \gamma)$ has a normal distribution with mean μ and variance σ^2 . The cdf of the three-parameter lognormal distribution is

$$F_x(x; \mu, \sigma, \gamma) = \Phi \left[\frac{\ln(x - \gamma) - \mu}{\sigma} \right] \quad (7)$$

For the three-parameter lognormal distribution defined in equation (7), the value of γ is given by the minimum population size value.

Estimation

Let $x_1, x_2, \dots, x_n$ be a sample of size n drawn from a PPS distribution. We assume that σ – parameter is given and we obtain it using the population of the smallest city. We will use the random variable Z defined by $Z = \log[X/\sigma]$ and its observed value by

$$z_i = \log(x_i / \sigma) \quad i = 1, 2, \dots, n$$

The log-likelihood function is given by

$$\log l(\lambda, \nu) = \sum_{i=1}^n \log f(x_i) = n \log \lambda + n \log \nu (\nu - 1) \sum_{i=1}^n \log z_i - \lambda \sum_{i=1}^n z_i - \sum_{i=1}^n \log x_i$$

where $f(x)$ is pdf defined in (5).

Maximum Likelihood Estimate of $\hat{\lambda}$ and $\hat{\nu}$

Taking partial derivatives with respect to λ and ν and equating then to zero the following normal equations are obtained.

$$\frac{\partial \log l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n z_i = 0 \quad (8)$$

$$\frac{\partial \log l}{\partial \nu} = \frac{n}{\nu} + \sum_{i=1}^n \log z_i - \lambda \sum_{i=1}^n z_i \log z_i = 0 \quad (9)$$

If λ is eliminated in Equations, (8) & (9) the equation in ν is obtained.

$$\frac{1}{\nu} + \frac{1}{n} \sum_{i=1}^n \log z_i - \frac{\sum_{i=1}^n z_i^\nu \log z_i}{\sum_{i=1}^n z_i^\nu} = 0 \quad (10)$$

The above equation can be solved using the Newton–Raphson method. The λ estimator

$$\hat{\lambda} = \left\{ \frac{1}{n} \sum_{i=1}^n z_i \right\}^{-1} \quad (11)$$

As already stated, more flexible models emerge when $\nu > 1$. The value of $\hat{\nu}$ is considered with the range $2.0 \leq \hat{\nu} \leq 2.5$.

Maximum Likelihood Estimation for Parameters μ and σ for Three-parameter Lognormal Distribution

The MLE for the parameters of μ and σ are given by

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \log(x_i - \gamma) \quad (12)$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n [\log(x_i - \gamma) - \hat{\mu}]^2 \quad (13)$$

Empirical Application to City Size

India has very rich source of information for urban studies. The census volumes, both at the national and state levels, provide a mine of information for rural and urban places for a period of 100 years. It is also main source of information for

the construction of city size distribution. The census periods covered are 1951, 1961, 1971, 1981, 1991 and 2001.

Urban population by size classification is based on the following:

Class-I	-	Population
I	-	Greater than 100,000
II	-	50,000 – 100,000
III	-	20,000 – 50,000
IV	-	10,000 – 20,000
V	-	5,000 – 10,000
VI	-	Less than 5,000

The number of cities / towns for each census year under six classes is given in the following [Table 1](#).

Table 1. Size Distribution of Cities and Towns in Kerala (1951-2001)

Census Year	> 100,000	50,000 – 100,000	20,000 – 50,000	10,000 – 20,000	5,000 – 10,000	< 5,000	Total
1951	4	3	10	21	6	1	45
1961	4	4	22	17	4	1	52
1971	5	8	32	11	3	1	60
1981	6	8	55	14	4	1	88
1991	9	17	69	34	10	1	140
2001	10	24	72	37	15	1	159

Data for Model Fitting

Some relevant information about the data sets used appears in following [Table 2](#). For each census year, the third column shows the size (number of people) of the smallest town we have considered. The fourth column shows the number of cities and towns fitted. The fifth column represents the percentage of the total Kerala cities / towns which have been considered. The sixth column shows the number of people who live in the cities and towns fitted and finally the seventh column the percentage of the total Kerala population that the number of inhabitants represents. For example, in 1951 we have considered 44 cities and towns with at least 3,098 people, which correspond to 97.78% of cities and towns and 99.79 of the total population of Kerala.

SIZE DISTRIBUTION OF CITIES IN KERALA

Table 2. Some relevant information about Kerala city size data sets used.

Census Year	Minimum town Size Considered	Town Considered		Population Considered	
		Number	% of Total	Number	% of Total
1951	3,098	44	97.78	14,85,347	99.79
1961	2,859	51	98.08	21,06,197	99.86
1971	4,750	59	98.33	30,68,436	99.84
1981	4,489	87	98.86	43,95,172	99.89
1991	4,820	139	99.28	72,57,261	99.94
2001	4,699	158	99.37	82,62,226	99.94

Fitted Models and Results

Three models were fitted and compared: classical Pareto distribution, three-parameter lognormal distribution and PPS distribution. The Pareto distribution was included for comparison purposes and it is known that this distribution is used to fit the upper tail of the distribution. The lognormal distribution was a classical distribution to fit a set of city size data. The PPS distribution was adjusted according to maximum likelihood method discussed in the [Estimation](#) section.

Akaike Information Criterion (AIC) for Pareto, Lognormal and PPS

For model identification Akaike (1974) suggested Akaike Information Criterion and the same is given by $AIC = 2\log l - 2d$ where $\log l$ is the likelihood of the model evaluated at the maximum likelihood estimates and d is the number of parameters.

The AIC is a measure of the goodness of fit of an estimated statistical model and a useful tool for model selection. In view of this we have to choose a model among the three models fitted which has the highest AIC. Parameter estimates and value of AIC statistics are given in the following [Table 3](#) for Pareto and lognormal distribution.

Table 3. Parameter estimates and value of AIC obtained from the fitting of the lognormal ($\hat{\mu}$ and $\hat{\sigma}$ parameters) and Pareto distribution ($\hat{\alpha}$ parameter) to the city size data in Kerala by maximum likelihood.

Census Year	N	$\hat{\mu}$	$\hat{\sigma}$	AIC - Lognormal	$\hat{\alpha}$	AIC - Pareto
1951	45	9.493	5.425	-1100.104	1.87	-1054.742
1961	52	3.557	1.005	-2783.805	2.147	-523.918
1971	60	4.304	0.948	-4960.471	1.829	-1048.969
1981	88	4.358	0.616	-1308.733	1.931	-2152.896
1991	140	2.546	1.266	-1271.844	0.829	-2521.164
2001	159	2.208	1.386	-1329.925	0.691	-2746.167

For all the data sets, the lognormal distribution presents a higher value of AIC statistics than the Pareto distribution. For example, in 2001 the value of AIC statistics is -1329.925 for the lognormal and -2746.167 for the Pareto distribution. In consequence, with these data sets the lognormal distribution is preferable to the Pareto distribution. This conclusion is consistent with the results obtained by Anderson and Ge (2005).

The results of PPS distribution appear in Table 4 for $2.0 < \hat{\nu} < 2.5$. In all the six considered census years, the distribution of PPS presents the highest values of the AIC statistics, in comparison with other two models. For example, in 2001 the AIC value is -688.802, higher than lognormal and Pareto AIC values. We can conclude that the distribution outperforms the classical Pareto and lognormal distribution in all the 6 data sets considered for the regional city size distribution.

Table 4. Parameter estimates and value of AIC obtained from the fitting of the PPS distribution ($\hat{\lambda}$ and $\hat{\nu}$ parameters) to the city size data in Kerala by maximum likelihood.

Census Year	N	$\hat{\lambda}$	$\hat{\nu}$	AIC -PPS
1951	45	0.813	2.1	-4749.968
1961	52	1.063	2.2	-535.435
1971	60	0.803	2.3	-580.813
1981	88	0.838	2.4	-1095.619
1991	140	2.675	2.5	-122.848
2001	159	2.332	2.6	-688.802

Conclusion

City size distribution data were analyzed using the PPS distribution developed by Sarabia and Prieto (2009). It provided a comparative flexible model for all range of a set of city size for six census periods. The lognormal distribution and Pareto distribution were also included comparison purpose because they are frequently used by urban researchers. The maximum likelihood estimate was the method used for the estimation of the parameters of lognormal Pareto, and PPS. Via AIC, it was noted that PPS distribution outperforms the fit provided by Pareto and lognormal distribution. This indicates that PPS is considered to be a good fit not only for country data but also for regional city size data.

References

- Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6), 716-723. doi:10.1109/TAC.1974.1100705
- Anderson, G., & Ge, Y. (2005). The size distribution of Chinese cities. *Regional Science and Urban Economics*, 35(6), 756-766. doi:10.1016/j.regsciurbeco.2005.01.003
- Auerbach, F. (1913). Das gesetz der bevölkerungskonzentration. *Petermanns Geographische Mitteilungen*, 59(1), 74-76.
- Brakman, S., Garretsen, H., Van Marrewijk, C., & Van Den Berg, M. (1999). The return of Zipf: Towards a further understanding of the rank-size distribution. *Journal of Regional Science*, 39(1), 182-213. doi:10.1111/1467-9787.00129
- Davis, D. R., & Weinstein, D. E. (2002). Bones, bombs, and break points: The geography of economic activity. *The American Economic Review*, 92(5), 1269-1289. doi:10.1257/000282802762024502
- Gabaix, X. (1999a). Zipf's law and the growth of cities. *The American Economic Review*, 89(2), 129-132. doi:10.1257/aer.89.2.129
- Gabaix, X. (1999b). Zipf's law for cities: An explanation. *The Quarterly Journal of Economics*, 114(3), 739-767. doi:10.1162/003355399556133
- Krugman, P. (1996). *The Self-Organizing Economy*. Cambridge, MA: Blackwell Publishing.
- Moura Jr., N. J., & Ribeiro, M. B. (2006). Zipf law for Brazilian cities. *Physica A: Statistical Mechanics and its Applications*, 367, 441-448. doi:10.1016/j.physa.2005.11.038

- Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, 57(1), 86-100. doi:[10.1016/j.jue.2004.09.002](https://doi.org/10.1016/j.jue.2004.09.002)
- Rosen, K. T., & Resnick, M. (1980). The size distribution cities: An examination of the Pareto law and primacy. *Journal of Urban Economics*, 8(2), 165-186. doi:[10.1016/0094-1190\(80\)90043-1](https://doi.org/10.1016/0094-1190(80)90043-1)
- Sarabia, J. M., & Prieto, F. (2009). The Pareto-positive stable distribution: A new descriptive model for city size data. *Physica A: Statistical Mechanics and its Applications*, 388(19), 4179-4191. doi:[10.1016/j.physa.2009.06.047](https://doi.org/10.1016/j.physa.2009.06.047)
- Simon, H. A. (1955). On a class of skewed distribution functions. *Biometrika*, 42(3/4), 425-440. doi:[10.2307/2333389](https://doi.org/10.2307/2333389)
- Soo, K. T. (2005). Zipf's law for cities: A cross country investigation. *Regional Science and Urban Economics*, 35(3), 239-263. doi:[10.1016/j.regsciurbeco.2004.04.004](https://doi.org/10.1016/j.regsciurbeco.2004.04.004)
- Subbarayan, A. (2009). The size distribution of cities in Tamilnadu (1901-2001). *International Journal of Agricultural and Statistical Sciences*, 5(2), 373-382.
- Zanette, D. H., & Manrubia, S. C. (1997). Role of intermittency in urban development: A model of large-scale city formation. *Physical Review Letters*, 79(3), 523-526. doi:[10.1103/PhysRevLett.79.523](https://doi.org/10.1103/PhysRevLett.79.523)
- Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

JMASM Algorithms and Code **A Percentile-Based Power Method in SAS: Simulating Multivariate Non-Normal Continuous Distributions**

Jennifer Koran
Southern Illinois University
Carbondale, Illinois

Todd C. Headrick
Southern Illinois University
Carbondale, Illinois

The conventional power method transformation is a moment-matching technique that simulates non-normal distributions with controlled measures of skew and kurtosis. The percentile-based power method is an alternative that uses the percentiles of a distribution in lieu of moments. This article presents a SAS/IML macro that implements the percentile-based power method.

Keywords: Monte Carlo, non-normal, percentiles, polynomial, SAS, simulation

---

Introduction

Fleishman (1978) introduced the power method, an elegant and convenient approach for simulating non-normal data by applying a third-order polynomial transformation to a normally distributed random variable. Multivariate extensions of this approach were later developed in Vale and Maurelli (1983) and Headrick and Sawilowsky (1999). A limitation of this conventional power method approach is that the third- and fourth-order moments of the distribution(s) must be available in order to implement the method (see Headrick & Sawilowsky, 2000).

Obtaining these higher-order moments is not a problem if individual data are available to the researcher. However, given privacy concerns and data restrictions, especially in education and health care, a researcher may desire to simulate data to match descriptive statistics available from publicly available reports. As third- and fourth-order moments are typically of little interest to the general public, these higher-order moments may not be part of publicly available reports.

---

*Dr. Koran is in the Section on Statistics and Measurement. Email her at: jkoran@siu.edu.
Dr. Headrick is in the Section on Statistics and Measurement, and is an Assistant Editor of this journal.*

To address this situation, Koran, Headrick, and Kuo (2015) introduced a multivariate power method that uses the percentiles of a distribution in place of moments. This approach relies upon the 10th, 25th, 50th, 75th, and 90th percentiles of each variable and can accommodate either Pearson or Spearman correlations between the variables. Further, Koran, Headrick, and Kuo (2015) showed that the percentile-based power method exhibits substantially lower relative bias than the conventional moment-based power method. Thus, the percentile-based power method is preferable to the conventional moment-based power method even when third- and fourth-order moments are available, as it mimics the characteristics of the original data more accurately. In light of the prior theoretical presentation of the percentile-based power method, this article focuses on the presentation of %simPPM, a SAS/IML (SAS Institute, 2013) macro that implements the percentile-based power method to produce the simulated data file.

The Percentile-based Power Method

This section presents the essential elements in applying the percentile-based power method to simulate non-normal univariate and multivariate data distributions. The third - order power method transformation is expressed as (Headrick, 2010, pp.12 - 13)

$$p(Z) = \sum_{i=1}^4 c_i Z^{i-1} \quad (1)$$

where $Z \sim \text{i.i.d. } N(0, 1)$ random variable with standard normal pdf ($\phi(z)$) and cdf ($\Phi(z)$). For the purposes contained herein, we assume that (1) is a strictly increasing, monotonic function. As such, an inverse function p^{-1} exists. Thus, the cdf of (1) is $F(p(z)) = (\Phi(z))$. Differentiating $F(p(z))$ with respect to z yields the pdf of (1), which is $f(p(z)) = \phi(z)/p'(z)$.

Univariate Non-normal Data Generation

The percentile-based power method begins with five percentile values ($q(x)_u$) that are used to produce estimates of the following location, scale, and shape parameters (Karian & Dudewicz, 2011, pp. 172-173)

$$\hat{\gamma}_1 = q(x)_{0.50} \quad (2)$$

$$\hat{\gamma}_2 = q(x)_{0.90} - q(x)_{0.10} \quad (3)$$

$$\hat{\gamma}_3 = \frac{q(x)_{0.50} - q(x)_{0.10}}{q(x)_{0.90} - q(x)_{0.50}} \quad (4)$$

$$\hat{\gamma}_4 = \frac{q(x)_{0.75} - q(x)_{0.25}}{\hat{\gamma}_2} \quad (5)$$

where (2)-(5) are the (i) median, (ii) inter-decile range, (iii) left-right tail-weight ratio (a skew function), and (iv) tail-weight factor (a kurtosis function), respectively.

The location, scale, and shape parameters computed from the percentile values are subsequently used, along with the standard normal constants of $z_{0.90} = 1.281\dots$ and $z_{0.75} = 0.6744\dots$, to solve for the following four coefficients (Koran, Headrick, & Kuo, 2015, Equations 47-50)

$$c_1 = \hat{\gamma}_1 \quad (6)$$

$$c_2 = \frac{\hat{\gamma}_2 (\hat{\gamma}_4 z_{0.90}^3 - z_{0.75}^3)}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3} \quad (7)$$

$$c_3 = \frac{\hat{\gamma}_2 (1 - \hat{\gamma}_3)}{2(1 + \hat{\gamma}_3) z_{0.90}^2} \quad (8)$$

$$c_4 = -\frac{\hat{\gamma}_2 (\hat{\gamma}_4 z_{0.90} - z_{0.75})}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3} \quad (9)$$

However, it is important to check that the solved values of the coefficients form a valid pdf under the conditions described earlier for the power method. In order to produce a valid non-normal pdf, the following criteria must be met (Koran, Headrick, & Kuo, 2015)

$$\begin{aligned} 0 < c_2 < 1 \\ 0 < c_4 < 0.608875 \\ c_3^2 - 3c_2c_4 < 0 \end{aligned} \quad (10)$$

Provided that the criteria in (10) are met, the power method coefficients from (6) - (9) are subsequently substituted into the third-order power method polynomial transformation in (1). The transformed values $p(Z)$ then represent a sample of size n from the univariate distribution defined by the original percentile values $(q(x)_u)$.

Multivariate Non-normal Data Generation

If it is desired to simulate multivariate data that will have a specified correlation structure following the transformation in (1), then standard normal random variables will first have to be simulated with an appropriate intermediate correlation structure. For each pair of variables Z_j, Z_k with $j \neq k$, the intermediate (Pearson) correlation r_{jk} can be determined by solving the following expression for r_{jk} (Headrick, 2010, Equation 4.34, p.114; Koran, Headrick, & Kuo, 2015, Equation 58)

$$\xi_{jk} = \frac{6}{\pi} \left\{ \left(\frac{n-2}{n-1} \right) \sin^{-1} \left(\frac{r_{jk}}{2} \right) + \left(\frac{1}{n-1} \right) \sin^{-1} (r_{jk}) \right\} \quad (11)$$

Where ξ_{jk} is fixed to a specified Spearman correlation.

In many cases a specified Pearson correlation is known instead of a specified Spearman correlation. This requires a different expression to be solved for the intermediate correlation. This new expression uses the following reduced form of Headrick's (2010) Equation 2.59 (p.30)

$$\begin{aligned} E[p(Z_j)p(Z_k)] = & (c_{j1}(c_{k1} + c_{k3}) + c_{k3}(c_{k1} + c_{k3})) \\ & + r_{jk}(c_{j2}c_{k2} + 3c_{j4}c_{k2} + 3c_{j2}c_{k4} + 9c_{j4}c_{k4}) \\ & + r_{jk}^2(2c_{j3}c_{k3}) + r_{jk}^3(6c_{j4}c_{k4}) \end{aligned} \quad (12)$$

where the fifth-order polynomial expression in Headrick (2010) has been reduced to the appropriate expression for a third-order polynomial and the coefficients $c_{j1} - c_{j4}$ and $c_{k1} - c_{k4}$ are the solved values of (6) - (9) for Z_j and Z_k , respectively.

PERCENTILE-BASED POWER METHOD IN SAS

The expression in (12) for using a specified Pearson correlation to solve for the intermediate (Pearson) correlation also uses the following expressions for computing the mean m and variance v from the percentile-based power method constant coefficients in (6) - (9) as

$$m_j = c_{j1} + c_{j3} \quad (13)$$

and

$$v_j = c_{j2}^2 + 2c_{j3}^3 + 6c_{j2}c_{j4} + 15c_{j4}^2 \quad (14)$$

Expressions analogous to (13) and (14) may be used for computing m_k and v_k , respectively. Thus, when a Pearson correlation is specified, the intermediate (Pearson) correlation r_{jk} can be found by substituting the expressions from (12), (13), and (14) into the following expression and solving for r_{jk}

$$\rho_{jk} = \frac{E[p(Z_j)p(Z_k)] - m_j m_k}{\sqrt{v_j v_k}} \quad (15)$$

Where ρ_{jk} is fixed to a specified Pearson correlation.

To apply the multivariate percentile-based power method, first compute the intermediate correlation(s) from either specified Spearman or Pearson correlation(s) in (11) or (15), respectively. Then, simulate standard normal random variables based on a Cholesky factorization of the intermediate correlation(s). Finally, apply the transformation in (1) to the individual simulated standard normal random variables. The resulting transformed values then represent a sample from the multivariate distribution with the specified correlation structure.

The Percentile-based Power Method in SAS/IML

The SAS/IML macro %simPPM implements the procedures described in the previous section. This %simPPM macro can be accessed by your program by including the following lines

```
filename simppm "directory of file simPPM";  
%include simppm(simPPM) / nosource2;
```

Calling %simPPM uses the following pieces of information: 1) the number of variables, 2) the file path and name of an external ASCII file with the percentiles, 3) the file path and name of an external ASCII file with the specified correlations (multivariate method only), 4) an indication of whether the specified correlations are Pearson or Spearman (1 for Pearson, 2 for Spearman; multivariate method only), 5) the desired sample size, 6) a random number seed (optional), and 7) the file path and name for the ASCII output file containing the simulated data.

Data in the percentiles and correlations files should be space delimited, and there cannot be any missing values. The percentiles are laid out such that there are five rows, the 10th percentile in the first row, the 25th percentile in the second row, the 50th percentile in the third row, the 75th percentile in the fourth row, and the 90th percentile in the fifth row. There are as many columns in the percentiles file as there are variables to be simulated. The correlations file has as many rows and columns as there are variables to be simulated, with the variables appearing in the same order as in the percentiles file, and the correlations should be arranged in a full symmetric matrix with ones on the diagonal. Examples of the layout for the percentiles and correlations files are shown in the [appendix](#) and explained further in the next section.

Examples

Univariate Example

Suppose we wish to simulate the Idaho Standards Achievement Test (ISAT) mathematics scale scores for 25 third grade students. The 2011 scale score to percentile rank conversion tables for the ISAT are publicly available ([Stoneberg, 2011](#)). We arranged the 10th, 25th, 50th, 75th, and 90th percentiles in the ASCII file ex1percentiles.txt for analysis as shown in the [appendix](#). With the ex1percentiles.txt file saved in the folder C:\SAS\, the %simPPM call for this example is:

```
%simPPM(1, C:\SAS\ex1percentiles.txt, , , 25, 54321, C:\SAS\ex1simdata.txt)
```

The "1" indicates that there is only one variable to be simulated. The "C:\SAS\ex1percentiles.txt" gives the file path and name of the ASCII file containing the percentiles. The next two arguments are left blank, as these are only needed for multivariate applications of the percentile-based power method. The "25" indicates that we would like a sample size of 25 for the simulated data.

PERCENTILE-BASED POWER METHOD IN SAS

The "54321" is a random number seed to use in generating the data. Alternately, the random number seed may be left blank. If the random number seed is not specified, SAS/IML will choose a random number seed, but the generated data will not be able to be replicated exactly because the random number seed will not be known. It is recommended that the user specify a random number seed. The "C:\SAS\ex1simdata.txt" indicates the file path and name of the ASCII file where the simulated data are to be stored. When the macro call is submitted successfully to SAS/IML, the percentile power method coefficients are produced as shown in the [appendix](#). As this is a small example with only 25 cases, the contents of ex1simdata.txt after submitting the macro call are also shown in the [appendix](#).

Multivariate Example

For the next example, percentiles and Pearson correlations were obtained from $n = 527$ respondents in the 2012 General Social Survey (Smith, Marsden, Hout, & Kim, 2013) who provided responses to two items. The first item was the respondent's age (AGE). The second item asked, "Approximately how much money or the cash equivalent of property have you contributed in each of the fields listed in the past 12 months? b. Education" (TOTEDUC) (Smith, Marsden, Hout, & Kim, 2013). Suppose we desired to simulate 1000 responses to these two survey items.

The 10th, 25th, 50th, 75th, and 90th percentiles were arranged for the AGE and TOTEDUC variables in the ASCII file ex2percentiles.txt for analysis as shown in the [appendix](#), and the specified Pearson correlation matrix for the AGE and TOTEDUC variables in the ASCII file ex2correlations.txt for analysis as shown in the [appendix](#). With the ex2percentiles.txt and ex2correlations.txt files saved in the folder C:\SAS\, the %simPPM call for this example is:

```
%simPPM(2, C:\SAS\ex2percentiles.txt, C:\SAS\ex2correlations.txt, 1, 1000,  
7654321, C:\SAS\ex2simdata.txt)
```

Figure 1. Distributions of 1000 cases of simulated data for two variables. Panel A depicts the variable AGE, and Panel B depicts TOTEDUC.

PERCENTILE-BASED POWER METHOD IN SAS

The "2" indicates that there are two variables to be simulated. The "C:\SAS\ex2percentiles.txt" gives the file path and name of the ASCII file containing the percentiles. The "C:\SAS\ex2correlations.txt" gives the file path and name of the ASCII file containing the specified correlation matrix. The "1" indicates that specified Pearson correlations are being provided in ex2correlations.txt (alternately a "2" here would indicate specified Spearman correlations). The "1000" indicates that we would like a sample size of 1000 for the simulated data. The "7654321" is a random number seed to use in generating the data. The "C:\SAS\ex2simdata.txt" indicates the file path and name of the ASCII file where the simulated data are to be stored. When the macro call is submitted successfully to SAS/IML, the percentile power method coefficients and intermediate correlation matrix are produced as shown in the [appendix](#).

The file ex2simdata.txt after submitting the macro call contains 1000 cases with two variables. The correlation between AGE and TOTEDUC is 0.0847 in the simulated data. Descriptive statistics and graphs summarizing the distribution of the simulated data are shown in [Figure 1](#).

Note that a small proportion of the cases of AGE are invalid (< 18 years), and a small proportion of the cases of TOTEDUC are invalid ($< \$0$). The distribution may be truncated and cases associated with these invalid values removed from the simulated data at the user's discretion.

Conclusion

The SAS/IML macro %simPPM simulates data to match univariate and multivariate non-normal distributions defined by percentiles and correlations. Unlike other power method approaches used for simulating non-normal data, %simPPM does not require the user to have direct access to conventional measures of skew and kurtosis, making this an ideal approach for simulating data to match distribution information available in public reports without direct access to individually identifiable data.

References

- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43(4), 521-532. doi:[10.1007/BF02293811](https://doi.org/10.1007/BF02293811)
- Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton, FL: CRC Press.
- Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, 64, 25-35. doi:[10.1007/BF02294317](https://doi.org/10.1007/BF02294317)
- Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, 25(4), 417-436. doi:[10.3102/10769986025004417](https://doi.org/10.3102/10769986025004417)
- Karian, Z. A., & Dudewicz, E. J. (2011). *Handbook of fitting statistical distributions with R*. Boca Raton FL: CRC Press.
- Koran, J., Headrick, T. C., & Kuo, T. C. (2015). Simulating univariate and multivariate nonnormal distributions through the method of percentiles. *Multivariate Behavioral Research*, 50(2), 216-232. doi:[10.1080/00273171.2014.963194](https://doi.org/10.1080/00273171.2014.963194)
- SAS Institute (2013). *Base SAS 9.4 procedures guide: Statistical procedures*. Cary, NC: author. <http://www.sas.com/>.
- Smith, T. W., Marsden, P. V., Hout, M. & Kim, J. (2013). General Social Survey 1972-2012 (SDA 4.0) [Data file]. Retrieved from <http://sda.berkeley.edu/archive.htm>.
- Stoneberg, B. (2011). Idaho Standards Achievement Tests Scale Score to Percentile Rank Conversion Tables Spring 2011. Boise, ID: Idaho State Department of Education Assessment Division.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, 48(3), 465-471. doi:[10.1007/BF02293687](https://doi.org/10.1007/BF02293687)

Appendix

Example Input and Output

Input and Output for Univariate Example

Input: The contents of the ASCII file ex1percentiles.txt are:

188.5
197.0
205.8
216.3
228.1

Output: The contents of the SAS output window are:

Percentile Power Method coefficients

C
205.8
13.869234
1.5221864
0.9625014

Output: The contents of ex1simdata.txt file are:

10.609710187
2.6051849323
75.05922126
29.634033258
62.861351473
80.326129351
99.236151941
83.577624847
100.16942799
81.800333938
52.766922741
110.68001386
43.654002567
74.415441414
52.073124328
77.869219855

43.866586956
126.84710015
80.419015418
58.74947208
112.64205437
30.900259549
96.584691093
105.65723356
44.858063646

Input and Output for Multivariate Example

Input: The contents of the ASCII file ex2percentiles.txt are:

27.176 15.500
33.667 25.350
41.444 90.339
48.901 180.737
59.500 600.529

Input: The contents of the ASCII file ex2correlations.txt are:

1 .10
.10 1

Output: The contents of the SAS output window are:

Percentile Power Method coefficients

C

41.444 90.339
10.78791 71.871855
1.1532084 132.53707
1.1102007 95.214843

Intermediate Pearson correlations

PI

1 0.1322937
0.1322937 1

JMASM Algorithms and Code **Nine Pseudo R^2 Indices for Binary Logistic Regression Models**

David A. Walker
Northern Illinois University
DeKalb, IL

Thomas J. Smith
Northern Illinois University
DeKalb, IL

This syntax program is an applied complement to Veall and Zimmermann (1994), Menard (2000), and Smith and McKenna (2013) and produces nine pseudo R^2 indices, not readily accessible in statistical software such as SPSS, which are used to describe the results from binary logistic regression analyses.

Keywords: Binary logistic regression, R^2 indices, SPSS, syntax

---

Introduction

The subsequent syntax-based software program (Walker & Smith, 2015) is intended to provide an application for users interested in various pseudo R^2 indices used to describe the results obtained from fitting binary logistic regression models, but not freely obtainable in the current SPSS operational format. In logistic regression, pseudo R^2 indices proffer an indication of model fit, and are similar to variance accounted for metrics affiliated with ordinary least-squares (OLS) regression models such as R^2 , R^2 adjusted, or eta squared. Although values of pseudo R^2 indices typically range from zero to unity, values for some indices can exceed 1.0.

The majority of the indices in the current program have been applied previously under various research conditions by Veall and Zimmermann (1994), Menard (2000), and Smith and McKenna (2013). This program is intended to be a software-based complement to these studies. Of the indices affiliated with the nine pseudo R^2 measures, only two are produced in SPSS: Cox and Snell's (1989) and Nagelkerke's (1991).

The Cox and Snell index is represented as

Dr. Walker is a Professor of Educational Research and Assessment. Email him at: dawalker@niu.edu.

$$R_{CS}^2 = 1 - \left(\frac{L(\text{Null})}{L(\text{Full})} \right)^{2/N}, \quad (1)$$

where $L(\text{Null})$ and $L(\text{Full})$ are the likelihood functions for the constant-only model and the model with the predictors, respectively, and N is the sample size. The Nagelkerke index, which is a “corrected” version of the Cox and Snell index in the sense that it constrains the index value so that it does not exceed 1.0, is expressed as

$$R_N^2 = \frac{1 - \left(\frac{L(\text{Null})}{L(\text{Full})} \right)^{2/N}}{1 - L(\text{Null})^{2/N}}, \quad (2)$$

where the rescaling is accomplished by dividing Cox and Snell’s index by its maximum possible value.

The remaining seven indices produced in the program are not produced in the default logistic regression output provided by SPSS. The McFadden (1974) and the McFadden adjusted metrics are stated as, respectively,

$$R_{MF}^2 = 1 - \frac{LL(\text{Full})}{LL(\text{Null})} \quad (3)$$

and

$$R_{MFA}^2 = 1 - \frac{LL(\text{Full}) - (K + 1)}{LL(\text{Null})}, \quad (4)$$

where $LL(\text{Full})$ is the log-likelihood value for the model containing the predictors and $LL(\text{Null})$ is the log-likelihood value for the constant-only model. The latter index “adjusts” (penalizes) for the number of predictors (K) in the model. The Aldrich and Nelson (1984) index is expressed as

$$R_{AN}^2 = \frac{2[LL(\text{Null}) - LL(\text{Full})]}{2[LL(\text{Null}) - LL(\text{Full})] + N}. \quad (5)$$

NINE PSEUDO R2 INDICES

The Veall and Zimmerman (1994) index, which is a “corrected” version of the Aldrich and Nelson index, is formulated as

$$R_{VZ}^2 = \frac{2[LL(Null) - LL(Full)]}{2[LL(Null) - LL(Full)] + N} \cdot \frac{2LL(Null) - N}{2LL(Full)}, \quad (6)$$

where $LL(Full)$ and $LL(Null)$ are as defined previously, and N is the sample size. The Sapra (2004) index is represented as

$$R_S^2 = 1 - \left(\frac{LL(Max) - LL(Full) + K + 1/2}{LL(Max) - LL(Null) + 1/2} \right), \quad (7)$$

where $LL(Null)$, $LL(Full)$, and K are as defined previously, and $LL(Max)$ indicates the maximum possible log-likelihood value for the saturated model (typically zero, for most data and models). Finally, the Estrella (1998) formulates both unadjusted and adjusted metrics, respectively,

$$R_E^2 = 1 - \left(\frac{LL(Full)}{LL(Null)} \right)^{\left(\frac{-2}{N} \right) LL(Null)} \quad (8)$$

and

$$R_{EA}^2 = 1 - \left(\frac{LL(Full) - K}{LL(Null)} \right)^{\left(\frac{-2}{N} \right) LL(Null)}, \quad (9)$$

where $LL(Null)$, $LL(Full)$, N , and K are as defined previously.

Program

The data set used in the current SPSS syntax program is a randomly sampled subset ($n = 200$) of the 1982 High School and Beyond data (Raudenbush & Bryk, 2002). Data were obtained from Acock (2008). To have the outcome variable (i.e., Write) fit the profile of a binary measure within a logistic regression model, the values were recoded, per the aforementioned data example, as a dichotomized variable

called “HonComp” with values $\geq 60 = 1$ or “Yes” and values $< 60 = 0$ or “No.” The full model also included three predictor variables: Female (0 = Male; 1 = Female), Read, and Science (note that the latter two variables were continuous in measurement).

The program is shown below. In the space between BEGIN DATA and END DATA, the user would insert the null model's -2 Log Likelihood (LL) value (LLNull), the full model's -2 LL value (LLFull), the sample size (N), and the number of predictors (K) in the model. The input values in the example from the program are, in this order, 231.289, 160.236, 200, 3, 0. Note that the last value place, LLMAX, has a suggested value that is frequently fixed at 0 and will be left as such in this example.

Pseudo R2 Program

*****

Copyright David A. Walker and Thomas J. Smith, 2015

Contact dawalker@niu.edu or tjsmith@niu.edu

Northern Illinois University, 204 Gabel, DeKalb, IL 60115

APA 6th Edition Citation

Walker, D. A., & Smith, T. J. (2015). Nine pseudo R2 indices for binary logistic regression models [Computer program]. DeKalb, IL: Authors.

*****

DATA LIST LIST / LLNull LLFull (2F9.3) N K (2F8.0) LLMAX.

*****

Between BEGIN DATA and END DATA below, put the Null Model's -2 Log Likelihood (LL)

value (LLNull), the Full Model's -2 LL value (LLFull), the sample size (N), and the number of

predictors (K) in the model Note: LLMAX, the last value in the data, is typically 0 rather than

always defaulted to 0

*****

BEGIN DATA

231.289 160.236 200 3 0

END DATA.

COMPUTE L0 = LLNull/-2.

COMPUTE L1 = LLFull/-2.

COMPUTE L2 = ((LLNull-LLFull)/2)+L0.

NINE PSEUDO R2 INDICES

```
COMPUTE CoxSnell = 1-EXP(-1*(LLNull-LLFull)/N).
COMPUTE Nagelkerke = CoxSnell/(1-EXP(-LLNull/N)).
COMPUTE McFadden = 1-(LLFull/LLNull).
COMPUTE McFaddenAdj = 1-((L1-(K+1))/L0).
COMPUTE Sapra = 1-((LLMax-LLFull+(K+1)/2)/(LLMax-LLNull+.5)).
COMPUTE AldrichNelson = ((LLNull-LLFull)/((LLNull-LLFull)+N)).
COMPUTE VeallZimmermann = AldrichNelson*((LLNull+N)/LLNull).
COMPUTE Estrella = 1-(L2/L0)**((-2/N)*L0).
COMPUTE EstrellaAdj = 1-((L2-K)/L0)**((-2/N)*L0).
EXECUTE.
FORMAT CoxSnell TO EstrellaAdj (F9.3).
VARIABLE LABELS CoxSnell 'Cox & Snell R2'/ Nagelkerke 'Nagelkerke R2'/
McFadden 'McFadden R2'/
VeallZimmermann 'Veall & Zimmermann R2'/Sapra 'Sapra R2'/McFaddenAdj
'McFadden Adjusted R2'/
AldrichNelson 'Aldrich & Nelson R2'/ Estrella 'Estrella R2'/ EstrellaAdj
'Estrella Adjusted R2'/.
REPORT FORMAT=LIST AUTOMATIC ALIGN (CENTER)
/VARIABLES= CoxSnell Nagelkerke McFadden McFaddenAdj AldrichNelson
Sapra VeallZimmermann Estrella EstrellaAdj
/TITLE "Pseudo R Squared Indices".
```

Results

As a simple, one-shot comparison, the values of pseudo R^2 obtained by applying the program to the High School and Beyond data, recorded in [Table 1](#), indicated that seven of the nine indices were much lower in value than the R^2 (0.522) or the R^2 adjusted (0.515) values computed from an OLS model using the same predictors as the logistic regression model and with the non-dichotomized, continuous outcome variable (i.e., Write). The aforementioned pseudo R^2 values ranged from a minimum of 0.262 (Aldrich and Nelson) to a maximum of 0.489 (Veall and Zimmermann). It should be noted, though, that the lower values of these indices compared to OLS R^2 values may reflect, in part, less precision in the outcome due to the dichotomization of the continuous dependent variable for use in logistic regression. Cohen (1983) remarks on of the cost of engaging in such research practices where, "...the cost in the degradation of measurement due to dichotomization is a loss of one-fifth to two-thirds of the variance, and a concomitant loss of power..." (p. 253).

The values of the Nagelkerke and the Veall and Zimmermann indices, both of which are “corrected” indices, were noticeably similar (i.e., 0.436 and 0.489, respectively) to the OLS R^2 values. These indices’ comparability in value to the OLS R^2 values was also found in Smith and McKenna (2013). Of interest is that the Smith and McKenna Monte Carlo simulation study, which included four continuous predictors, and the current findings indicate, potentially, that the Veall and Zimmermann pseudo R^2 index’s highly favorable comparisons to OLS R^2 values may signify a robust nature within this index toward countering the full effect of dichotomizing an outcome variable in binary logistic regression modeling.

Table 1. Pseudo R^2 results

Cox-Snell	Nagelkerke	McFadden	McFadden Adj.	Aldritch-Nelson
0.299	0.436	0.307	0.273	0.626
Sapra	Veall-Zimmerman	Estrella	Estrella Adj.	
0.314	.0489	0.346	0.317	

References

- Acock, A. C. (2008). *A gentle introduction to Stata* (2nd ed.). College Station, TX: Stata Press.
- Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Beverly Hills, CA: Sage.
- Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, 7, 249-253. doi: [10.1177/014662168300700301](https://doi.org/10.1177/014662168300700301)
- Cox, D. R., & Snell, E. J. (1989). *The analysis of binary data* (2nd ed.). London, UK: Chapman and Hall.
- Estrella, A. (1998). A new measure of fit for equations with dichotomous dependent variables. *Journal of Business & Economic Statistics*, 16(2), 198-205. doi: [10.1080/07350015.1998.10524753](https://doi.org/10.1080/07350015.1998.10524753)
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 104-142). New York, NY: Academic Press.
- Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, 54(1), 17-24. doi: [10.1080/00031305.2000.10474502](https://doi.org/10.1080/00031305.2000.10474502)

NINE PSEUDO R² INDICES

Nagelkerke, N. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, 78(3), 691-692. doi: [10.2307/2337038](https://doi.org/10.2307/2337038)

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Sapra, S. K. (2004). Comment by Sapra and reply. *The American Statistician*, 58(1), 90. doi: [10.1198/0003130042917](https://doi.org/10.1198/0003130042917)

Smith, T. J., & McKenna, C. M. (2013). A comparison of logistic regression pseudo R^2 indices. *Multiple Linear Regression Viewpoints*, 39(2), 17-26. Retrieved from http://www.glmj.org/archives/articles/Smith_v39n2.pdf

Veall, M. R., & Zimmermann, K. F. (1994). Evaluating pseudo- R^2 's for binary probit models. *Quality and Quantity*, 28(2), 151-164. doi: [10.1007/BF01102759](https://doi.org/10.1007/BF01102759)

Walker, D. A., & Smith, T. J. (2015). Nine pseudo R^2 indices for binary logistic regression models [Computer program]. DeKalb, IL: Authors.

JMASM Algorithms and Code Simple Response Surface Methodology Using RSREG (SAS)

Wan Muhammad Amir
University of Science, Malaysia
Kelantan, Malaysia

Kasypi Mokhtar
University of Malaysia
Terengganu, Malaysia

Hanafi A. Rahim
University of Malaysia
Terengganu, Malaysia

Mohamad Shafiq
University of Science, Malaysia
Kelantan, Malaysia

Nor Azlida Aleng
University of Malaysia
Terengganu, Malaysia

Zalila Ali
University of Science, Malaysia
Penang, Malaysia

Response surface methodology (RSM) can be used when the response variable, y , is influenced by several variables, x 's. When treatments take the form of quantitative values, then the true relationship between response variables and independent variables might be known. Examples are given in SAS.

Keywords: Multiple linear regression, response surface methodology

---

Introduction

Response surface methodology (RSM) was introduced by Box and Wilson (1951). The RSM explores the relationships between several explanatory variables (X) and one or more response variables (Y). The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response through linear models and second-degree polynomials. This model is only an approximation, but it is easy to apply even when little is known about the process. According to Montgomery (2005), RSM is a statistical technique that useful for modelling and analysis of problems in which a response of interest is influenced by some variables and the objective is to optimize the response variable, Y . The general form of RSM can be expressed as $y = f(x_1, x_2, x_3, \dots, x_n) + \varepsilon$ or $y = f'(\mathbf{x})\boldsymbol{\beta} + e$ where $\mathbf{x} = (x_1, x_2, \dots, x_k)'$, $f(\mathbf{x})$ is a vector function of p elements that consists of powers and cross-products of powers of $x_1, x_2, \dots, x_k$ up to a certain degree, denoted by $d (\geq 1)$, $\boldsymbol{\beta}$ is a vector of p unknown coefficients referred to as parameters, and e is an experimental error term.

Dr. Amir bin W Ahmad is an Associate Professor of Biostatistics. Email him at: wmamir@usm.my. Mohamad Shafiq Bin Mohd Ibrahim is a postgraduate student in the School of Dental Sciences. Email him at: shafiqmat786@gmail.com.

SIMPLE RESPONSE SURFACE METHODOLOGY USING SAS

Two important models are commonly used in RSM. These are the first-degree model ($d = 1$),

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon ,$$

and the second-degree model ($d = 2$),

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \beta_{ij} x_i x_j + \sum_{i=1}^k \beta_{ii} x_i^2 + \varepsilon .$$

The relationship between y and $x_1, x_2, \dots, x_k$ can be used to predict response values for given settings of the control variables. Also, the significance of the factor whose levels are presented by $x_1, x_2, \dots, x_k$ can be determined and the optimum setting of $x_1, x_2, \dots, x_k$ over a certain region of interest can be identified. RSMs are designs and models for searching for the optimum response through linear and second-degree polynomial models. When there is more than one response, then it is important to find the compromise optimum that does not optimize only one response (Myers, Khuri, & Carter; Oehlert, 2000; Ngo, 2012). In general, the response surface can be visualized graphically; this graph is very helpful in order to see the shape of a response surface.

As an example, consider the function of y (Triglycerides) plotted against the levels of x_1 (BMI) and x_2 (F7) as shown in Figure 1 (Ahmad, Shafiq, Halim, & Aleng, 2014).

Figure 1. Surface plot of Trig vs F7, BMI, left; Contour plot of Trig vs F7, BMI, right

Materials and Methods

The relationship between response variable and independent variable is quite difficult to determine. The model parameters can be estimated most effectively if proper experimental design is used to collect the data. The relationship between response and independent variable is determined by a mathematical model called a regression model. There are two models involved, the first-order model and second-order model.

First-order model (method of least square)

In general, a first-order model takes the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

with two independent variables and, with n independent variables,

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_q x_q + \varepsilon_i, \quad i = 1, 2, \dots, n$$

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \varepsilon$$

where $(\beta_0, \beta_1, \beta_2, \dots, \beta_q)$ are regression coefficients, $(x_0, x_1, x_2, \dots, x_q)$ are independent or predictor variables, ε is random error, and y is a dependent, or response, variable.

Second-order model (method of least square)

With two independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \varepsilon$$

In general, a second-order model or second-degree polynomial (with n independent variables) expression takes the following form:

$$y = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^k \beta_{ij} x_i x_j + \sum_{i=1}^k \beta_{ii} x_i^2 + \varepsilon.$$

The second degree polynomial is flexible because it can take a variety of function forms. The β_0 , β_i , and β_{ij} are constant and ε is a term of error or residual between the observed and calculated value.

Case study I: First-order design

Suppose A and B denote the natural variables temperature (ζ_1), pressure ratio (ζ_2), and genuineness (Y). Then the transformation of these natural variables to coded variables is

$$C = \frac{X_1 + 220}{5}, \quad D = \frac{X_2 - 1.2}{0.1}.$$

The calculation of the coded variables is shown in Table 2.

The relationship between the response variable y and independent variable $\mathbf{x}$ is usually unknown. In general, the lower order polynomial model is used to describe the response surface, f . Results from Table 3 indicate that the linear models ($p = 0.015$) were statistically significant, suggesting that this model adequately fits the data.

The first stage of the model was fitted to the data by the method of least squares. The regression model is given by

$$Y = 84.1 + 0.850C + 0.250D.$$

Table 1. Process data for fitting, first-order model (natural variables) (Montgomery, 1984)

Temperature (ζ_1) X_1	Pressure Ratio (ζ_2) X_2	Genuineness (Y)
-225	1.1	82.8
-225	1.3	83.5
-215	1.1	84.7
-215	1.3	85.0
-220	1.2	84.1
-220	1.2	84.5
-220	1.2	83.9
-220	1.2	84.3

Table 2. Process data for fitting, first-order model (coded variables)

Temperature (ζ_1) C	Pressure Ration (ζ_2) D	Genuineness (Y)
-1	-1	82.8
-1	1	83.5
1	-1	84.7
1	1	85.0
0	0	84.1
0	0	84.5
0	0	83.9
0	0	84.3

Table 3. Analysis of variance for genuineness (Y)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	4	3.26	3.26	0.81500	12.23	0.034
Linear	2	3.14	3.14	1.57000	23.55	0.015
Square	1	0.08	0.08	0.08000	1.20	0.353
Interaction	1	0.04	0.04	0.04000	0.60	0.495
Residual Error	3	0.20	0.20	0.06667		
Pure Error	3	0.20	0.20	0.06667		
Total	7	3.46				

Table 4. Estimated regression coefficients for Y

Predictor	Coef	SE Coef	T	P
Constant	84.10	0.0894	940.27	0.000
C	0.85	0.1265	6.72	0.001
D	0.25	0.1265	1.98	0.105

Table 5. Analysis of variance (ANOVA)

Source	DF	SS	MS	F	P
Regression	2	3.14	1.570	24.53	0.003
Residual Error	5	0.32	0.064		
Total	7	3.46			

SIMPLE RESPONSE SURFACE METHODOLOGY USING SAS

The residual plots do not indicate any problem with the model. A normal distribution with a mean of 1.246×10^{-14} and a standard deviation of 0.1690 appears to fit our sample data fairly well. The plotted points form a reasonably straight line. In our case, the residuals bounce randomly around the 0 line (residual vs. predicted value). This suggests that the assumption that the relationship is linear is reasonable. A higher *R*-squared value of 1 indicates how well the data fits the model and also indicates a better model.

Figure 2. Residual plot for Y

Figure 3. Contour and surface plot for genuineness Y vs. temperature X_1 and pressure ratio X_2 .

The counter and surface plots indicate that the highest value of Genuineness is obtained when temperature is high and pressure ratio level is high. This area appears at the upper right corner of the plot.

Performing response surface analysis using the SAS RSREG procedure

```

Data predict ;
input x1 x2 Y;
cards;
-225  1.1  82.8
-225  1.3  83.5
-215  1.1  84.7
-220  1.2  84.1
-220  1.2  84.5
-220  1.2  83.9
-220  1.2  84.3
;
/*plots=(surface)*/
ods graphics on;
proc rsreg data=predict plots=(surface);
model y=x1 x2/lackfit;
run;

```


SIMPLE RESPONSE SURFACE METHODOLOGY USING SAS

```
ods graphics off;

/*surface(3D)*/
ods graphics on;
proc rsreg data=predict plots=surface(3D);
model y=x1 x2/lackfit;
run;
ods graphics off;

/*Plot all*/
ods graphics on;
proc rsreg data=predict plots=all;
model y=x1 x2/lackfit;
run;
ods graphics off;
```

Case study II: Second-order design

Let A and B denote the natural variables reading (ζ_1) X_1 , reading (ζ_2) X_2 , and response (Y). Then the transformation of these natural variables to coded variables is

$$C = \frac{X_1 - 95}{5}, \quad D = \frac{X_2 - 165}{5}$$

The calculation of the coded variables is shown in [Table 7](#).

The relationship between the response variable Y and independent X_1 and X_2 is usually unknown. In general, the lower order polynomial models are used to describe the response surface, f. Because the linear model suggested that a higher model is needed to adequately model the response surface, the full quadratic model is fitted. For the full quadratic model ([Table 8](#)), the p-value for lack of fit is 0.089, suggesting that this model adequately fits the data.

The second stage of the model was also fitted to the data by the method of least squares. We get the following model in coded variables by using the regression method. The fitted regression model is given by

$$Y = 81.00 + C + 0.5D - 1.5C^2$$

Table 6. Process data for fitting, second-order model (natural variables) (Montgomery, 1984)

Reading (ζ_1) X_1	Reading (ζ_2) X_2	Response (Y)
90	160	78
90	170	79
100	160	80
100	170	81
95	165	81
95	165	82
95	165	82
95	165	80
95	165	80

Table 7. Process data for fitting, second-order model (coded variables)

Reading (ζ_1) C	Reading (ζ_2) D	Response (Y)
-1	-1	78
-1	1	79
1	-1	80
1	1	81
0	0	81
0	0	82
0	0	82
0	0	80
0	0	80

Table 8. Analysis of variance for genuineness (Y)

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Regression	4	10	10	2.5	2.5	0.198
Linear	2	5	5	2.5	2.5	0.198
Square	1	5	5	5.0	5.0	0.089
Interaction	1	0	0	0.0	0.0	1.000
Residual Error	4	4	4	1.0		
Pure Error	4	4	4	1.0		
Total	8	14				

SIMPLE RESPONSE SURFACE METHODOLOGY USING SAS

Table 9. Estimated regression coefficients for Y

Predictor	Coef	SE Coef	T	P
Constant	81.0	0.4472	181.122	0.000
C	1.0	0.5000	2.000	0.116
D	0.5	0.5000	1.000	0.374
C ²	-1.5	0.6708	-2.236	0.089

Figure 4. Residual plot for Y

Figure 5. Contour and surface plot for response Y vs. reading (X_1) and reading (X_2)

The residual plots do not indicate any problem with the model. In our case, the residuals bounce randomly around the 0 line (residual vs predicted value). This suggests that the assumption that the relationship is linear is reasonable. How well the estimated model fits the data can be measured by the value of R^2 . The R^2 lies in the interval $[0, 1]$. A higher R-squared value of 0.99 indicated how well the data fit the model and also indicates a better model.

The counter and surface plots indicate that the highest value of response Y is obtained when the reading X_2 is high and the reading X_1 is in the range of 92 to 98. This area appears at the right corner of the plot. In addition, we can see the shape of the of the response surface and get a general idea of the response Y at various settings of reading X_1 and reading X_2 .

Response surface methodology are design and models for working with continuous treatments when finding the optima or describing the response is the goal (Oehlert, 2000).

Performing response surface analysis using the SAS RSREG procedure

```
Data predict ;
input x1 x2 Y;
cards;
```

90	160	78
90	170	79

SIMPLE RESPONSE SURFACE METHODOLOGY USING SAS

100	160	80
100	170	81
95	165	81
95	165	82
95	165	80
95	165	80

```
;
/*plots=(surface)*/
ods graphics on;
proc rsreg data=predict plots=(surface);
model y=x1 x2/lackfit;
run;
ods graphics off;

/*surface(3D)*/
ods graphics on;
proc rsreg data=predict plots=surface(3D);
model y=x1 x2/lackfit;
run;
ods graphics off;

/*Plot all*/
ods graphics on;
proc rsreg data=predict plots=all;
model y=x1 x2/lackfit;
run;
ods graphics off;
```

Summary and Conclusion

Factorial designs with the RSMs provided a preliminary idea of the dependent variables with two independent variables by plotting a contour and surface response function. The factorial designs are widely used in experiments when the curvature in the response surface is concerned. Actually, it is easier to understand the behavior of the data by using graphing and canonical analysis. This response surface method reveals the finding with more explicitly due to the surface plot performance. It provides the comprehensive information and also give the general idea of dependent variables at the various setting of two independent variables. After a

proper design is conducted, the response surface analysis can be made by any statistical computer software as such SAS and then statistical analyses can be applied to draw the appropriate conclusions from the study.

References

- Ahmad, W. M. A. W., Shafiq, M., Halim, N., & Aleng, N. A. (2014). A study of triglycerides level in three distinct phases of human blood pressure: A case study from previous projects. *Applied Mathematical Sciences*, (8)46, 2289-2305. doi: [10.12988/ams.2014.42145](https://doi.org/10.12988/ams.2014.42145)
- Box, G. E. P. & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society. Series B (methodological)* 13(1), 1-45. Available from <http://www.jstor.org/stable/2983966>
- Montgomery, D. C. (1984). *Design and analysis of experiments* (2nd ed). New York: Wiley.
- Montgomery, D. C. (2005). *Design and analysis of experiments: Response surface method and designs* (6th ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Myers, R. H., Khuri, A. I., & Carter, W. H., Jr. (1989). Response surface methodology: 1966-1988. *Technometrics* 31(2), 137-153. doi: [10.1080/00401706.1989.10488509](https://doi.org/10.1080/00401706.1989.10488509)
- Ngo, T. H. D. (2012). The steps to follow in a multiple regression analysis. *Proceedings of the SAS Global Forum 2012 Conference (paper 333-2012)*. Cary, NC: SAS Institute Inc. Retrieved from <http://support.sas.com/resources/papers/proceedings12/333-2012.pdf>
- Oehlert, G. W. (2000). *A first course in design and analysis of experiments*. New York, NY: W. H. Freeman.

JMASM Algorithms and Code Confidence Intervals for Kendall's Tau with Small Samples

David A. Walker
Northern Illinois University
DeKalb, IL

A syntax program, not readily expedient in statistical software such as SPSS, is provided for an application of confidence interval estimates with Kendall's tau-b for small samples.

Keywords: Kendall's tau, confidence intervals, SPSS, syntax

Introduction

This syntax program (Walker, 2015) is intended to provide an application, not readily available, for users in SPSS who are interested in a range of confidence interval (CI) estimates with Kendall's tau-b (τ) for small samples. Statistically, Kendall's tau-b is a non-parametric, correlational method typically employed with ordinal x and y measures. Tau's application within social science research has been predominantly in correlational meta-analysis studies and also as a component of experimental research (Cooper & Hedges, 1994; Gilpin, 1993). Furthermore, tau has been highlighted as a proxy for Pearson's product-moment correlation (r) in research situations where sample sizes are small (Rupinski & Dunlap, 1996). Tau is expressed as (Helsel & Hirsch, 1995):

$$\tau = \frac{C - D}{N(N-1)/2} \quad (1)$$

where C = number of concordant pairs, D = number of discordant pairs, and N = sample size.

Dr. Walker is a Professor of Educational Research and Assessment. Email him at: dawalker@niu.edu.

According to Bonett and Wright (2000), “interval estimation may be more appropriate in applications where the magnitude of a correlation is of primary interest” (p. 23). Certainly, there has been prominence afforded in the literature to this issue that confidence intervals need to be reported with point estimates, such as tau, to supplement and progress the interpretation of outcomes (American Psychological Association, 2010; Levin & Robinson, 2003).

Further, Long and Cliff (1997), supposing a bivariate normal population, found that tau performed reasonably well with small samples > 10 and < 25 , where samples ≥ 25 could be conducted with a Pearson r correlation because of x and y 's sampling from a bivariate normal distribution (Kendall, 1949). Lower and upper bound confidence intervals for tau can be represented through a series of steps such that Fisher's z -transformation is used for “normalizing the sampling distribution of τ ” (Long & Cliff, 1997, p. 35). Specifically,

$$\tau_z = 0.5 \ln \left(\frac{1+\tau}{1-\tau} \right) . \quad (2)$$

Per Fieller, Hartley, and Pearson (1957), the standard error of τ_z [$SE(\tau_z)$] is used with the desired unit normal critical value to construct the lower and upper confidence bounds for τ_z :

$$\tau_{z,\text{lower}} = \tau_z - z_{1-\alpha} SE(\tau_z) \quad (3)$$

and

$$\tau_{z,\text{upper}} = \tau_z + z_{1-\alpha} SE(\tau_z) , \quad (4)$$

where $\tau_{z,\text{lower}}$ and $\tau_{z,\text{upper}}$ are the lower and upper bounds, respectively; $z_{1-\alpha}$ is the unit normal critical z value for specified level of confidence $1 - \alpha$; and

$$SE(\tau_z) = \sqrt{\frac{0.437}{N-4}} \quad (5)$$

The values of the previously-mentioned lower and upper confidence bounds for τ_z are then transformed (Fisher, 1925) due to tau's ability “...to generalize from the sample to the population correlation for any monotonic transformation of bivariate normal variables” (Bonett & Wright, 2000, p. 24). Specifically,

CONFIDENCE INTERVALS FOR KENDALL'S TAU

$$\tau_{\text{lower}} = \frac{\exp(2\tau_{\text{lower}}) - 1}{\exp(2\tau_{\text{lower}}) + 1} \quad (6)$$

and

$$\tau_{\text{upper}} = \frac{\exp(2\tau_{\text{upper}}) - 1}{\exp(2\tau_{\text{upper}}) + 1} \quad (7)$$

Confidence Intervals for Kendall's Tau with Small Samples Program

The independent SPSS syntax platform allows the user to merely run the program, which yields a comprehensive list of tau values ranging from 0.99 to -0.99, in increments of 0.01, for small samples extending from 11 to 24, per the Long and Cliff (1997) recommendation. The confidence intervals presented in conjunction with the tau values are programmed at 80%, 90%, 95%, and 99%, where Cohen (1990, p. 1310) observed "I don't think that we should routinely use 95% intervals: Our interests are often better served by more tolerant 80% intervals." Further, Tukey (1960) added that the application of confidence intervals was of paramount importance,

Probably the greatest ultimate importance, among all types of statistical procedures we now know, belongs to confidence procedures which, by making interval estimates, attempt to reach as strong conclusions as are reasonable by pointing out, not single likely values, but rather whole classes (intervals, regions, etc.) of possible values, so chosen that there can be high confidence that the "true" value is somewhere among them. (p. 429)

It should be noted that this program produces an exceedingly large table (i.e., Table 1 is over 60 pages) and; therefore, only a sample of the results are shown below. The selection of results display an assortment of magnitude and directionality with tau and small samples from 11 to 24. The array of example tau values include positive, large tau ranging from 0.90 to 0.99; positive, moderate values from 0.40 to 0.50; and negative, small values extending from -0.20 to -0.10, all with accompanying CIs.

DAVID A. WALKER

*****

Copyright David A. Walker, 2015

Contact dawalker@niu.edu

Northern Illinois University, 204D Gabel, DeKalb, IL 60115

APA 6th Edition Citation

Walker, D. A. (2015). Confidence intervals for Kendall's tau with small samples [Computer program]. DeKalb, IL: Author.

*****.

INPUT PROGRAM.

LOOP #CASE = -99 TO 99.

LOOP #N = 11 TO 24.

COMPUTE T = #CASE*.01.

COMPUTE N = #N.

END CASE.

END LOOP.

END LOOP.

END FILE.

END INPUT PROGRAM.

EXECUTE.

SORT CASES BY

T (D) .

SELECT IF NOT (T = .00).

COMPUTE FISHERZ = .5*LN((1+T)/(1-T)).

COMPUTE SEZ = SQRT(.437/(N-4)).

COMPUTE CRITICAL80 = ABS(IDF.NORMAL((1-.80)/2,0,1)).

COMPUTE FISHERZL80 = FISHERZ - 1.28*SEZ.

COMPUTE FISHERZU80 = FISHERZ + 1.28*SEZ.

COMPUTE TL80 = (EXP(2*FISHERZL80) - 1)/(EXP(2*FISHERZL80) + 1).

COMPUTE TU80 = (EXP(2*FISHERZU80) - 1)/(EXP(2*FISHERZU80) + 1).

COMPUTE CRITICAL90 = ABS(IDF.NORMAL((1-.90)/2,0,1)).

COMPUTE FISHERZL90 = FISHERZ - 1.645*SEZ.

COMPUTE FISHERZU90 = FISHERZ + 1.645*SEZ.

COMPUTE TL90 = (EXP(2*FISHERZL90) - 1)/(EXP(2*FISHERZL90) + 1).

COMPUTE TU90 = (EXP(2*FISHERZU90) - 1)/(EXP(2*FISHERZU90) + 1).

COMPUTE CRITICAL95 = ABS(IDF.NORMAL((1-.95)/2,0,1)).

COMPUTE FISHERZL95 = FISHERZ - 1.96*SEZ.

COMPUTE FISHERZU95 = FISHERZ + 1.96*SEZ.

COMPUTE TL95 = (EXP(2*FISHERZL95) - 1)/(EXP(2*FISHERZL95) + 1).

COMPUTE TU95 = (EXP(2*FISHERZU95) - 1)/(EXP(2*FISHERZU95) + 1).

COMPUTE CRITICAL99 = ABS(IDF.NORMAL((1-.99)/2,0,1)).

COMPUTE FISHERZL99 = FISHERZ - 2.58*SEZ.

COMPUTE FISHERZU99 = FISHERZ + 2.58*SEZ.

COMPUTE TL99 = (EXP(2*FISHERZL99) - 1)/(EXP(2*FISHERZL99) + 1).

COMPUTE TU99 = (EXP(2*FISHERZU99) - 1)/(EXP(2*FISHERZU99) + 1).

EXECUTE.

FORMAT N (F8.0).

FORMAT T (F9.2).

FORMAT FISHERZ TO TU99 (F9.3).

CONFIDENCE INTERVALS FOR KENDALL'S TAU

```
VARIABLE LABELS T 'Tau' /TU90 'Upper 90% CI'/TL90 'Lower 90% CI'/TU95
'Upper 95% CI'/TL95 'Lower 95% CI'/TU80 'Upper 80% CI'/TL80 'Lower 80%
CI'/TU99 'Upper 99% CI'/TL99 'Lower 99% CI'.
REPORT FORMAT=LIST AUTOMATIC ALIGN (CENTER)
/VARIABLES= N T TL99 TL95 TL90 TL80 TU80 TU90 TU95 TU99
/TITLE "Sample Size, Kendall's Tau, and 80%, 90%, 95%, 99% Confidence
Intervals".
```

Table 1. Sample size, Kendall's tau, and confidence intervals

N	Tau	L 99% CI	L 95% CI	L 90% CI	L 80% CI	U 80% CI	U 90% CI	U 95% CI	U 99% CI
11	0.99	0.964	0.974	0.977	0.981	0.995	0.996	0.996	0.997
12	0.99	0.967	0.975	0.979	0.982	0.994	0.995	0.996	0.997
13	0.99	0.969	0.976	0.979	0.982	0.994	0.995	0.996	0.997
14	0.99	0.971	0.977	0.980	0.983	0.994	0.995	0.996	0.997
15	0.99	0.972	0.978	0.981	0.983	0.994	0.995	0.995	0.996
16	0.99	0.973	0.979	0.981	0.984	0.994	0.995	0.995	0.996
17	0.99	0.974	0.980	0.982	0.984	0.994	0.995	0.995	0.996
18	0.99	0.975	0.980	0.982	0.984	0.994	0.994	0.995	0.996
19	0.99	0.976	0.981	0.983	0.985	0.994	0.994	0.995	0.996
20	0.99	0.977	0.981	0.983	0.985	0.993	0.994	0.995	0.996
21	0.99	0.977	0.981	0.983	0.985	0.993	0.994	0.995	0.996
22	0.99	0.978	0.982	0.983	0.985	0.993	0.994	0.995	0.996
23	0.99	0.978	0.982	0.984	0.985	0.993	0.994	0.994	0.995
24	0.99	0.979	0.982	0.984	0.985	0.993	0.994	0.994	0.995
11	0.98	0.929	0.948	0.955	0.962	0.989	0.991	0.992	0.994
12	0.98	0.935	0.951	0.957	0.964	0.989	0.991	0.992	0.994
13	0.98	0.939	0.953	0.959	0.965	0.989	0.990	0.992	0.994
14	0.98	0.942	0.955	0.961	0.966	0.988	0.990	0.991	0.993
15	0.98	0.945	0.957	0.962	0.967	0.988	0.990	0.991	0.993
16	0.98	0.947	0.958	0.963	0.968	0.988	0.989	0.990	0.992
17	0.98	0.949	0.959	0.964	0.968	0.987	0.989	0.990	0.992
18	0.98	0.951	0.960	0.965	0.969	0.987	0.989	0.990	0.992
19	0.98	0.952	0.961	0.965	0.969	0.987	0.989	0.990	0.992
20	0.98	0.954	0.962	0.966	0.970	0.987	0.988	0.989	0.991
21	0.98	0.955	0.963	0.966	0.970	0.987	0.988	0.989	0.991
22	0.98	0.956	0.963	0.967	0.970	0.987	0.988	0.989	0.991
23	0.98	0.957	0.964	0.967	0.971	0.986	0.988	0.989	0.991
24	0.98	0.958	0.965	0.968	0.971	0.986	0.988	0.989	0.991
11	0.97	0.895	0.922	0.933	0.944	0.984	0.987	0.989	0.992
12	0.97	0.903	0.927	0.936	0.946	0.983	0.986	0.988	0.991
13	0.97	0.909	0.930	0.939	0.948	0.983	0.985	0.987	0.990
14	0.97	0.914	0.933	0.941	0.949	0.982	0.985	0.987	0.990
15	0.97	0.918	0.936	0.943	0.951	0.982	0.984	0.986	0.989
16	0.97	0.922	0.938	0.945	0.952	0.981	0.984	0.986	0.989
17	0.97	0.925	0.939	0.946	0.952	0.981	0.983	0.985	0.988

DAVID A. WALKER

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
18	0.97	0.927	0.941	0.947	0.953	0.981	0.983	0.985	0.988
19	0.97	0.929	0.942	0.948	0.954	0.981	0.983	0.985	0.987
20	0.97	0.931	0.943	0.949	0.955	0.980	0.982	0.984	0.987
21	0.97	0.933	0.944	0.950	0.955	0.980	0.982	0.984	0.987
22	0.97	0.934	0.945	0.950	0.956	0.980	0.982	0.984	0.986
23	0.97	0.936	0.946	0.951	0.956	0.980	0.982	0.983	0.986
24	0.97	0.937	0.947	0.952	0.957	0.979	0.981	0.983	0.986
11	0.96	0.862	0.897	0.911	0.926	0.979	0.982	0.985	0.989
12	0.96	0.872	0.903	0.916	0.928	0.978	0.981	0.984	0.988
13	0.96	0.880	0.908	0.919	0.931	0.977	0.980	0.983	0.987
14	0.96	0.887	0.911	0.922	0.933	0.976	0.980	0.982	0.986
15	0.96	0.892	0.915	0.924	0.934	0.976	0.979	0.981	0.986
16	0.96	0.896	0.917	0.926	0.936	0.975	0.978	0.981	0.985
17	0.96	0.900	0.920	0.928	0.937	0.975	0.978	0.980	0.984
18	0.96	0.903	0.922	0.930	0.938	0.974	0.977	0.980	0.984
19	0.96	0.906	0.923	0.931	0.939	0.974	0.977	0.979	0.983
20	0.96	0.909	0.925	0.932	0.940	0.974	0.977	0.979	0.983
21	0.96	0.911	0.926	0.933	0.940	0.973	0.976	0.978	0.982
22	0.96	0.913	0.928	0.934	0.941	0.973	0.976	0.978	0.982
23	0.96	0.915	0.929	0.935	0.942	0.973	0.976	0.978	0.982
24	0.96	0.916	0.930	0.936	0.942	0.972	0.975	0.977	0.981
11	0.95	0.830	0.872	0.890	0.907	0.973	0.978	0.981	0.986
12	0.95	0.842	0.880	0.895	0.911	0.972	0.977	0.980	0.985
13	0.95	0.852	0.885	0.899	0.914	0.971	0.975	0.979	0.984
14	0.95	0.860	0.890	0.903	0.916	0.970	0.975	0.978	0.983
15	0.95	0.866	0.894	0.906	0.918	0.970	0.974	0.977	0.982
16	0.95	0.872	0.897	0.908	0.920	0.969	0.973	0.976	0.981
17	0.95	0.876	0.900	0.910	0.921	0.968	0.972	0.975	0.980
18	0.95	0.880	0.902	0.912	0.923	0.968	0.972	0.975	0.980
19	0.95	0.883	0.905	0.914	0.924	0.967	0.971	0.974	0.979
20	0.95	0.887	0.907	0.915	0.925	0.967	0.971	0.974	0.978
21	0.95	0.889	0.908	0.917	0.926	0.967	0.970	0.973	0.978
22	0.95	0.892	0.910	0.918	0.926	0.966	0.970	0.973	0.977
23	0.95	0.894	0.911	0.919	0.927	0.966	0.969	0.972	0.977
24	0.95	0.896	0.912	0.920	0.928	0.965	0.969	0.972	0.976
11	0.94	0.798	0.848	0.869	0.889	0.968	0.973	0.977	0.983
12	0.94	0.813	0.856	0.875	0.893	0.967	0.972	0.976	0.982
13	0.94	0.824	0.863	0.880	0.897	0.965	0.970	0.974	0.980
14	0.94	0.833	0.869	0.884	0.900	0.964	0.969	0.973	0.979
15	0.94	0.841	0.873	0.888	0.902	0.964	0.968	0.972	0.978
16	0.94	0.847	0.877	0.890	0.904	0.963	0.968	0.971	0.977
18	0.94	0.857	0.884	0.895	0.907	0.961	0.966	0.970	0.975
19	0.94	0.861	0.886	0.897	0.909	0.961	0.965	0.969	0.975
20	0.94	0.865	0.888	0.899	0.910	0.960	0.965	0.968	0.974
21	0.94	0.868	0.890	0.900	0.911	0.960	0.964	0.968	0.973

CONFIDENCE INTERVALS FOR KENDALL'S TAU

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
22	0.94	0.871	0.892	0.902	0.912	0.959	0.964	0.967	0.973
23	0.94	0.873	0.894	0.903	0.913	0.959	0.963	0.966	0.972
24	0.94	0.876	0.895	0.904	0.914	0.959	0.963	0.966	0.972
11	0.93	0.767	0.824	0.848	0.871	0.962	0.969	0.973	0.980
12	0.93	0.784	0.834	0.855	0.876	0.961	0.967	0.971	0.979
13	0.93	0.797	0.842	0.861	0.880	0.960	0.965	0.970	0.977
14	0.93	0.807	0.848	0.865	0.883	0.958	0.964	0.969	0.976
15	0.93	0.816	0.853	0.869	0.886	0.957	0.963	0.967	0.974
16	0.93	0.823	0.858	0.873	0.888	0.956	0.962	0.966	0.973
17	0.93	0.829	0.861	0.876	0.890	0.956	0.961	0.965	0.972
18	0.93	0.834	0.865	0.878	0.892	0.955	0.960	0.964	0.971
19	0.93	0.839	0.868	0.880	0.894	0.954	0.959	0.964	0.970
20	0.93	0.843	0.870	0.882	0.895	0.954	0.959	0.963	0.970
21	0.93	0.847	0.873	0.884	0.896	0.953	0.958	0.962	0.969
22	0.93	0.850	0.875	0.886	0.897	0.952	0.957	0.961	0.968
23	0.93	0.853	0.877	0.887	0.898	0.952	0.957	0.961	0.967
24	0.93	0.856	0.878	0.889	0.899	0.952	0.956	0.960	0.967
11	0.92	0.737	0.800	0.827	0.854	0.957	0.964	0.969	0.977
12	0.92	0.756	0.811	0.835	0.859	0.955	0.962	0.967	0.975
13	0.92	0.770	0.820	0.842	0.864	0.954	0.960	0.965	0.974
14	0.92	0.782	0.827	0.847	0.867	0.952	0.959	0.964	0.972
15	0.92	0.791	0.833	0.851	0.870	0.951	0.958	0.963	0.971
16	0.92	0.799	0.838	0.855	0.873	0.950	0.956	0.961	0.969
17	0.92	0.806	0.842	0.858	0.875	0.949	0.955	0.960	0.968
18	0.92	0.812	0.846	0.861	0.877	0.948	0.954	0.959	0.967
19	0.92	0.817	0.850	0.864	0.879	0.948	0.954	0.958	0.966
20	0.92	0.822	0.852	0.866	0.880	0.947	0.953	0.957	0.965
21	0.92	0.826	0.855	0.868	0.882	0.946	0.952	0.957	0.964
22	0.92	0.830	0.857	0.870	0.883	0.946	0.951	0.956	0.963
23	0.92	0.833	0.860	0.872	0.884	0.945	0.951	0.955	0.963
24	0.92	0.836	0.862	0.873	0.885	0.945	0.950	0.954	0.962
11	0.91	0.708	0.777	0.806	0.836	0.951	0.959	0.965	0.974
12	0.91	0.728	0.789	0.815	0.842	0.950	0.957	0.963	0.972
13	0.91	0.744	0.799	0.823	0.847	0.948	0.955	0.961	0.970
14	0.91	0.757	0.807	0.829	0.851	0.946	0.954	0.959	0.968
15	0.91	0.767	0.813	0.834	0.854	0.945	0.952	0.958	0.967
16	0.91	0.776	0.819	0.838	0.857	0.944	0.951	0.956	0.965
17	0.91	0.784	0.824	0.841	0.860	0.943	0.950	0.955	0.964
18	0.91	0.790	0.828	0.845	0.862	0.942	0.949	0.954	0.963
19	0.91	0.796	0.832	0.847	0.864	0.941	0.948	0.953	0.962
20	0.91	0.801	0.835	0.850	0.866	0.940	0.947	0.952	0.961
22	0.91	0.809	0.840	0.854	0.869	0.939	0.945	0.950	0.959
23	0.91	0.813	0.843	0.856	0.870	0.938	0.944	0.949	0.958
24	0.91	0.816	0.845	0.858	0.871	0.937	0.944	0.949	0.957
11	0.90	0.679	0.754	0.786	0.819	0.946	0.955	0.961	0.971

DAVID A. WALKER

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
12	0.90	0.701	0.767	0.796	0.825	0.944	0.952	0.959	0.969
13	0.90	0.718	0.778	0.804	0.831	0.942	0.950	0.957	0.967
14	0.90	0.732	0.787	0.810	0.835	0.940	0.948	0.955	0.965
15	0.90	0.743	0.794	0.816	0.839	0.939	0.947	0.953	0.963
16	0.90	0.753	0.800	0.820	0.842	0.937	0.945	0.951	0.961
17	0.90	0.761	0.805	0.824	0.845	0.936	0.944	0.950	0.960
18	0.90	0.768	0.810	0.828	0.847	0.935	0.943	0.949	0.959
19	0.90	0.775	0.814	0.831	0.849	0.934	0.942	0.948	0.957
20	0.90	0.780	0.817	0.834	0.851	0.933	0.941	0.946	0.956
21	0.90	0.785	0.820	0.836	0.853	0.933	0.940	0.945	0.955
22	0.90	0.790	0.823	0.838	0.855	0.932	0.939	0.944	0.954
23	0.90	0.794	0.826	0.840	0.856	0.931	0.938	0.944	0.953
24	0.90	0.797	0.828	0.842	0.857	0.930	0.937	0.943	0.952
11	0.50	-0.095	0.060	0.137	0.226	0.701	0.744	0.778	0.832
12	0.50	-0.054	0.091	0.163	0.245	0.690	0.732	0.765	0.819
13	0.50	-0.019	0.117	0.185	0.261	0.681	0.722	0.754	0.807
14	0.50	0.010	0.139	0.203	0.275	0.673	0.713	0.744	0.796
15	0.50	0.035	0.157	0.218	0.286	0.667	0.705	0.735	0.787
16	0.50	0.057	0.174	0.231	0.296	0.660	0.698	0.727	0.779
17	0.50	0.076	0.188	0.243	0.305	0.655	0.692	0.720	0.771
18	0.50	0.093	0.200	0.253	0.312	0.650	0.686	0.714	0.764
19	0.50	0.109	0.212	0.262	0.319	0.646	0.681	0.708	0.757
20	0.50	0.122	0.222	0.271	0.325	0.642	0.676	0.703	0.751
21	0.50	0.135	0.231	0.278	0.331	0.638	0.671	0.698	0.746
22	0.50	0.146	0.239	0.285	0.336	0.634	0.667	0.694	0.740
23	0.50	0.157	0.247	0.291	0.341	0.631	0.663	0.689	0.735
24	0.50	0.166	0.254	0.297	0.345	0.628	0.660	0.685	0.731
11	0.49	-0.108	0.046	0.124	0.213	0.694	0.738	0.772	0.828
12	0.49	-0.067	0.078	0.150	0.233	0.683	0.726	0.759	0.814
13	0.49	-0.032	0.104	0.172	0.249	0.674	0.716	0.748	0.802
14	0.49	-0.003	0.126	0.190	0.262	0.666	0.706	0.738	0.791
15	0.49	0.022	0.144	0.205	0.274	0.659	0.698	0.729	0.782
16	0.49	0.044	0.161	0.219	0.284	0.653	0.691	0.721	0.773
17	0.49	0.063	0.175	0.230	0.293	0.647	0.685	0.714	0.765
18	0.49	0.080	0.188	0.241	0.300	0.642	0.679	0.708	0.758
19	0.49	0.095	0.199	0.250	0.307	0.638	0.673	0.702	0.752
20	0.49	0.109	0.209	0.258	0.314	0.634	0.668	0.696	0.745
21	0.49	0.122	0.218	0.266	0.319	0.630	0.664	0.691	0.740
22	0.49	0.133	0.227	0.273	0.324	0.626	0.660	0.687	0.734
23	0.49	0.144	0.234	0.279	0.329	0.623	0.656	0.682	0.729
24	0.49	0.153	0.241	0.285	0.334	0.620	0.652	0.678	0.725
12	0.48	-0.080	0.065	0.138	0.220	0.676	0.720	0.754	0.810
13	0.48	-0.045	0.091	0.159	0.236	0.667	0.709	0.742	0.797
14	0.48	-0.016	0.113	0.177	0.250	0.659	0.700	0.732	0.787
15	0.48	0.009	0.132	0.193	0.262	0.652	0.692	0.723	0.777

CONFIDENCE INTERVALS FOR KENDALL'S TAU

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
16	0.48	0.031	0.148	0.206	0.272	0.645	0.684	0.715	0.768
17	0.48	0.050	0.162	0.218	0.281	0.640	0.678	0.708	0.760
18	0.48	0.067	0.175	0.228	0.288	0.635	0.672	0.701	0.753
19	0.48	0.082	0.186	0.238	0.295	0.630	0.666	0.695	0.746
20	0.48	0.096	0.196	0.246	0.302	0.626	0.661	0.689	0.739
21	0.48	0.109	0.206	0.254	0.307	0.622	0.657	0.684	0.734
22	0.48	0.120	0.214	0.261	0.313	0.618	0.652	0.680	0.728
23	0.48	0.131	0.222	0.267	0.317	0.615	0.648	0.675	0.723
24	0.48	0.141	0.229	0.273	0.322	0.612	0.645	0.671	0.718
11	0.47	-0.134	0.020	0.099	0.188	0.680	0.726	0.762	0.819
12	0.47	-0.093	0.052	0.125	0.208	0.669	0.714	0.748	0.805
13	0.47	-0.058	0.078	0.147	0.224	0.660	0.703	0.736	0.793
14	0.47	-0.029	0.100	0.165	0.238	0.651	0.693	0.726	0.782
15	0.47	-0.004	0.119	0.180	0.250	0.644	0.685	0.717	0.772
16	0.47	0.018	0.135	0.194	0.260	0.638	0.677	0.708	0.763
17	0.47	0.037	0.150	0.205	0.269	0.632	0.671	0.701	0.754
18	0.47	0.054	0.162	0.216	0.277	0.627	0.664	0.694	0.747
19	0.47	0.070	0.174	0.225	0.284	0.622	0.659	0.688	0.740
20	0.47	0.083	0.184	0.234	0.290	0.618	0.654	0.683	0.734
21	0.47	0.096	0.193	0.241	0.296	0.614	0.649	0.677	0.728
22	0.47	0.108	0.202	0.248	0.301	0.610	0.645	0.673	0.722
23	0.47	0.118	0.210	0.255	0.306	0.607	0.641	0.668	0.717
24	0.47	0.128	0.217	0.261	0.310	0.604	0.637	0.664	0.712
11	0.46	-0.146	0.008	0.086	0.176	0.674	0.720	0.756	0.815
12	0.46	-0.105	0.039	0.112	0.196	0.662	0.707	0.742	0.801
13	0.46	-0.071	0.065	0.134	0.212	0.652	0.696	0.730	0.788
14	0.46	-0.042	0.087	0.152	0.226	0.644	0.686	0.720	0.777
15	0.46	-0.017	0.106	0.168	0.238	0.637	0.678	0.710	0.766
16	0.46	0.005	0.123	0.181	0.248	0.630	0.670	0.702	0.757
17	0.46	0.024	0.137	0.193	0.257	0.624	0.663	0.695	0.749
18	0.46	0.041	0.150	0.204	0.265	0.619	0.657	0.688	0.741
19	0.46	0.057	0.161	0.213	0.272	0.614	0.652	0.681	0.734
20	0.46	0.071	0.172	0.222	0.278	0.610	0.646	0.676	0.728
21	0.46	0.083	0.181	0.229	0.284	0.606	0.642	0.670	0.722
22	0.46	0.095	0.190	0.236	0.289	0.602	0.637	0.666	0.716
23	0.46	0.106	0.197	0.243	0.294	0.599	0.633	0.661	0.711
24	0.46	0.115	0.205	0.249	0.299	0.596	0.629	0.657	0.706
11	0.45	-0.159	-0.005	0.074	0.163	0.667	0.714	0.751	0.811
12	0.45	-0.118	0.027	0.100	0.183	0.655	0.701	0.737	0.796
13	0.45	-0.084	0.053	0.122	0.200	0.645	0.690	0.724	0.783
14	0.45	-0.055	0.075	0.140	0.214	0.637	0.680	0.714	0.772
16	0.45	-0.008	0.110	0.169	0.236	0.622	0.663	0.696	0.752
17	0.45	0.012	0.125	0.181	0.245	0.617	0.656	0.688	0.743
18	0.45	0.029	0.138	0.192	0.253	0.611	0.650	0.681	0.735
19	0.45	0.044	0.149	0.201	0.260	0.606	0.644	0.675	0.728

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
20	0.45	0.058	0.159	0.210	0.267	0.602	0.639	0.669	0.722
21	0.45	0.071	0.169	0.217	0.272	0.598	0.634	0.663	0.715
22	0.45	0.083	0.177	0.224	0.278	0.594	0.630	0.658	0.710
23	0.45	0.093	0.185	0.231	0.283	0.591	0.626	0.654	0.704
24	0.45	0.103	0.193	0.237	0.287	0.588	0.622	0.649	0.699
11	0.44	-0.171	-0.017	0.061	0.151	0.660	0.708	0.745	0.806
12	0.44	-0.130	0.014	0.088	0.171	0.648	0.695	0.731	0.791
13	0.44	-0.096	0.040	0.109	0.188	0.638	0.683	0.718	0.778
14	0.44	-0.067	0.062	0.128	0.202	0.629	0.673	0.707	0.766
15	0.44	-0.042	0.081	0.143	0.214	0.621	0.664	0.698	0.756
16	0.44	-0.020	0.098	0.157	0.224	0.615	0.656	0.689	0.746
17	0.44	-0.001	0.112	0.169	0.233	0.609	0.649	0.681	0.738
18	0.44	0.016	0.125	0.180	0.241	0.603	0.643	0.674	0.730
19	0.44	0.032	0.137	0.189	0.248	0.598	0.637	0.668	0.722
20	0.44	0.046	0.147	0.198	0.255	0.594	0.632	0.662	0.716
21	0.44	0.059	0.157	0.206	0.261	0.590	0.627	0.656	0.709
22	0.44	0.070	0.165	0.213	0.266	0.586	0.622	0.651	0.704
23	0.44	0.081	0.173	0.219	0.271	0.583	0.618	0.647	0.698
24	0.44	0.091	0.181	0.225	0.276	0.579	0.614	0.642	0.693
11	0.43	-0.183	-0.030	0.049	0.139	0.653	0.702	0.740	0.802
12	0.43	-0.142	0.002	0.075	0.159	0.641	0.688	0.725	0.787
13	0.43	-0.108	0.028	0.097	0.176	0.630	0.676	0.712	0.773
14	0.43	-0.079	0.050	0.115	0.190	0.622	0.666	0.701	0.761
15	0.43	-0.054	0.069	0.131	0.202	0.614	0.657	0.691	0.751
16	0.43	-0.032	0.086	0.145	0.212	0.607	0.649	0.683	0.741
17	0.43	-0.013	0.100	0.157	0.221	0.601	0.642	0.675	0.732
18	0.43	0.004	0.113	0.168	0.230	0.595	0.635	0.667	0.724
19	0.43	0.020	0.125	0.177	0.237	0.590	0.630	0.661	0.716
20	0.43	0.034	0.135	0.186	0.243	0.586	0.624	0.655	0.710
21	0.43	0.046	0.145	0.194	0.249	0.582	0.619	0.649	0.703
22	0.43	0.058	0.153	0.201	0.255	0.578	0.615	0.644	0.697
23	0.43	0.069	0.161	0.207	0.260	0.574	0.610	0.639	0.692
24	0.43	0.078	0.169	0.213	0.264	0.571	0.606	0.635	0.686
11	0.42	-0.194	-0.042	0.037	0.127	0.645	0.696	0.734	0.798
12	0.42	-0.154	-0.010	0.063	0.147	0.633	0.682	0.719	0.782
13	0.42	-0.120	0.016	0.085	0.164	0.623	0.670	0.706	0.768
14	0.42	-0.091	0.038	0.103	0.178	0.614	0.659	0.695	0.756
15	0.42	-0.066	0.057	0.119	0.190	0.606	0.650	0.685	0.745
16	0.42	-0.045	0.074	0.133	0.201	0.599	0.642	0.676	0.735
17	0.42	-0.025	0.088	0.145	0.210	0.593	0.635	0.668	0.726
18	0.42	-0.008	0.101	0.156	0.218	0.587	0.628	0.661	0.718
20	0.42	0.021	0.123	0.174	0.232	0.578	0.617	0.648	0.703
21	0.42	0.034	0.133	0.182	0.238	0.574	0.612	0.642	0.697
22	0.42	0.046	0.141	0.189	0.243	0.570	0.607	0.637	0.691
23	0.42	0.056	0.149	0.196	0.248	0.566	0.603	0.632	0.685

CONFIDENCE INTERVALS FOR KENDALL'S TAU

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
24	0.42	0.066	0.157	0.202	0.253	0.563	0.599	0.628	0.680
11	0.41	-0.206	-0.054	0.025	0.115	0.638	0.689	0.728	0.793
12	0.41	-0.166	-0.022	0.051	0.136	0.626	0.675	0.713	0.777
13	0.41	-0.132	0.004	0.073	0.152	0.615	0.663	0.700	0.763
14	0.41	-0.103	0.026	0.091	0.166	0.606	0.652	0.689	0.751
15	0.41	-0.078	0.045	0.107	0.179	0.598	0.643	0.678	0.740
16	0.41	-0.057	0.062	0.121	0.189	0.591	0.635	0.669	0.730
17	0.41	-0.037	0.076	0.133	0.198	0.585	0.627	0.661	0.720
18	0.41	-0.020	0.089	0.144	0.206	0.580	0.621	0.654	0.712
19	0.41	-0.005	0.101	0.154	0.214	0.574	0.615	0.647	0.704
20	0.41	0.009	0.111	0.162	0.220	0.570	0.609	0.641	0.697
21	0.41	0.022	0.121	0.170	0.226	0.565	0.604	0.635	0.691
22	0.41	0.034	0.129	0.177	0.232	0.562	0.599	0.630	0.685
23	0.41	0.044	0.137	0.184	0.237	0.558	0.595	0.625	0.679
24	0.41	0.054	0.145	0.190	0.242	0.554	0.591	0.620	0.673
11	0.40	-0.217	-0.066	0.013	0.103	0.631	0.683	0.723	0.789
12	0.40	-0.177	-0.034	0.039	0.124	0.619	0.669	0.707	0.773
13	0.40	-0.144	-0.008	0.061	0.141	0.608	0.656	0.694	0.758
14	0.40	-0.115	0.014	0.080	0.155	0.599	0.645	0.682	0.746
15	0.40	-0.090	0.033	0.095	0.167	0.591	0.636	0.672	0.734
16	0.40	-0.069	0.050	0.109	0.177	0.584	0.628	0.663	0.724
17	0.40	-0.049	0.064	0.121	0.187	0.577	0.620	0.654	0.715
18	0.40	-0.032	0.077	0.132	0.195	0.572	0.613	0.647	0.706
19	0.40	-0.017	0.089	0.142	0.202	0.566	0.607	0.640	0.698
20	0.40	-0.003	0.099	0.151	0.209	0.562	0.602	0.634	0.691
21	0.40	0.010	0.109	0.159	0.215	0.557	0.596	0.628	0.684
22	0.40	0.022	0.118	0.166	0.221	0.553	0.591	0.622	0.678
23	0.40	0.032	0.126	0.172	0.226	0.550	0.587	0.617	0.672
24	0.40	0.042	0.133	0.179	0.230	0.546	0.583	0.613	0.667
11	-0.10	-0.632	-0.530	-0.471	-0.397	0.216	0.301	0.371	0.496
12	-0.10	-0.606	-0.507	-0.450	-0.380	0.196	0.277	0.343	0.464
13	-0.10	-0.584	-0.487	-0.432	-0.365	0.180	0.256	0.320	0.437
14	-0.10	-0.565	-0.470	-0.417	-0.352	0.166	0.239	0.300	0.413
15	-0.10	-0.547	-0.455	-0.404	-0.341	0.154	0.224	0.282	0.392
16	-0.10	-0.532	-0.442	-0.392	-0.332	0.143	0.210	0.267	0.373
17	-0.10	-0.518	-0.430	-0.382	-0.323	0.134	0.199	0.253	0.356
18	-0.10	-0.505	-0.419	-0.372	-0.315	0.125	0.188	0.241	0.341
19	-0.10	-0.494	-0.409	-0.364	-0.308	0.118	0.179	0.230	0.328
20	-0.10	-0.483	-0.401	-0.356	-0.302	0.111	0.170	0.220	0.315
21	-0.10	-0.473	-0.392	-0.349	-0.296	0.105	0.162	0.211	0.303
22	-0.10	-0.464	-0.385	-0.342	-0.291	0.099	0.155	0.202	0.293
24	-0.10	-0.448	-0.371	-0.331	-0.282	0.089	0.142	0.187	0.274
11	-0.11	-0.638	-0.537	-0.479	-0.406	0.206	0.292	0.362	0.489
12	-0.11	-0.613	-0.514	-0.458	-0.388	0.187	0.267	0.334	0.456
13	-0.11	-0.591	-0.495	-0.441	-0.374	0.170	0.247	0.311	0.429

DAVID A. WALKER

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
14	-0.11	-0.572	-0.478	-0.425	-0.361	0.156	0.229	0.291	0.404
15	-0.11	-0.554	-0.463	-0.412	-0.350	0.144	0.214	0.273	0.383
16	-0.11	-0.539	-0.450	-0.401	-0.341	0.133	0.201	0.258	0.364
17	-0.11	-0.525	-0.438	-0.390	-0.332	0.124	0.189	0.244	0.347
18	-0.11	-0.513	-0.427	-0.381	-0.324	0.115	0.178	0.232	0.332
19	-0.11	-0.501	-0.418	-0.372	-0.318	0.108	0.169	0.220	0.318
20	-0.11	-0.491	-0.409	-0.365	-0.311	0.101	0.160	0.210	0.306
21	-0.11	-0.481	-0.401	-0.358	-0.306	0.094	0.152	0.201	0.294
22	-0.11	-0.472	-0.393	-0.351	-0.300	0.089	0.145	0.193	0.284
23	-0.11	-0.463	-0.387	-0.345	-0.295	0.083	0.138	0.185	0.274
24	-0.11	-0.456	-0.380	-0.340	-0.291	0.079	0.132	0.177	0.264
11	-0.12	-0.644	-0.544	-0.487	-0.414	0.197	0.283	0.353	0.481
12	-0.12	-0.619	-0.522	-0.466	-0.397	0.177	0.258	0.325	0.448
13	-0.12	-0.597	-0.502	-0.449	-0.382	0.160	0.237	0.302	0.420
14	-0.12	-0.578	-0.486	-0.434	-0.370	0.146	0.220	0.281	0.396
15	-0.12	-0.561	-0.471	-0.421	-0.359	0.134	0.204	0.264	0.375
16	-0.12	-0.546	-0.458	-0.409	-0.349	0.123	0.191	0.248	0.356
17	-0.12	-0.532	-0.446	-0.399	-0.341	0.114	0.179	0.234	0.339
18	-0.12	-0.520	-0.436	-0.390	-0.333	0.105	0.168	0.222	0.323
19	-0.12	-0.509	-0.426	-0.381	-0.327	0.098	0.159	0.211	0.309
20	-0.12	-0.498	-0.417	-0.373	-0.320	0.091	0.150	0.201	0.297
21	-0.12	-0.489	-0.409	-0.366	-0.315	0.084	0.142	0.191	0.285
22	-0.12	-0.480	-0.402	-0.360	-0.310	0.079	0.135	0.183	0.274
23	-0.12	-0.471	-0.395	-0.354	-0.305	0.073	0.128	0.175	0.264
24	-0.12	-0.464	-0.389	-0.349	-0.300	0.069	0.122	0.168	0.255
11	-0.13	-0.650	-0.551	-0.494	-0.422	0.187	0.273	0.344	0.473
12	-0.13	-0.625	-0.529	-0.474	-0.405	0.167	0.248	0.316	0.440
13	-0.13	-0.604	-0.510	-0.457	-0.391	0.150	0.228	0.292	0.412
14	-0.13	-0.585	-0.493	-0.442	-0.379	0.136	0.210	0.272	0.387
15	-0.13	-0.568	-0.479	-0.429	-0.368	0.124	0.195	0.254	0.366
16	-0.13	-0.553	-0.466	-0.417	-0.358	0.113	0.181	0.239	0.347
17	-0.13	-0.540	-0.454	-0.407	-0.350	0.104	0.169	0.225	0.330
18	-0.13	-0.527	-0.444	-0.398	-0.342	0.095	0.159	0.212	0.314
19	-0.13	-0.516	-0.434	-0.390	-0.336	0.088	0.149	0.201	0.300
20	-0.13	-0.506	-0.426	-0.382	-0.330	0.081	0.140	0.191	0.287
21	-0.13	-0.496	-0.418	-0.375	-0.324	0.074	0.132	0.181	0.276
22	-0.13	-0.487	-0.410	-0.369	-0.319	0.069	0.125	0.173	0.265
23	-0.13	-0.479	-0.404	-0.363	-0.314	0.063	0.118	0.165	0.255
24	-0.13	-0.472	-0.397	-0.357	-0.309	0.058	0.112	0.158	0.246
11	-0.14	-0.656	-0.558	-0.502	-0.431	0.177	0.264	0.335	0.465
12	-0.14	-0.632	-0.536	-0.482	-0.414	0.157	0.239	0.307	0.432
14	-0.14	-0.592	-0.501	-0.450	-0.387	0.126	0.200	0.263	0.379
15	-0.14	-0.575	-0.487	-0.437	-0.377	0.114	0.185	0.245	0.357
16	-0.14	-0.560	-0.474	-0.426	-0.367	0.103	0.171	0.229	0.338
17	-0.14	-0.547	-0.462	-0.416	-0.359	0.093	0.159	0.215	0.320

CONFIDENCE INTERVALS FOR KENDALL'S TAU

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
18	-0.14	-0.535	-0.452	-0.407	-0.351	0.085	0.149	0.203	0.305
19	-0.14	-0.524	-0.443	-0.398	-0.345	0.077	0.139	0.191	0.291
20	-0.14	-0.513	-0.434	-0.391	-0.339	0.070	0.130	0.181	0.278
21	-0.14	-0.504	-0.426	-0.384	-0.333	0.064	0.122	0.172	0.266
22	-0.14	-0.495	-0.419	-0.378	-0.328	0.058	0.115	0.163	0.255
23	-0.14	-0.487	-0.412	-0.372	-0.323	0.053	0.108	0.155	0.245
24	-0.14	-0.479	-0.406	-0.366	-0.319	0.048	0.102	0.148	0.236
11	-0.15	-0.662	-0.565	-0.510	-0.439	0.167	0.254	0.326	0.457
12	-0.15	-0.638	-0.544	-0.490	-0.422	0.147	0.229	0.298	0.423
13	-0.15	-0.617	-0.525	-0.473	-0.408	0.130	0.208	0.274	0.395
14	-0.15	-0.598	-0.509	-0.458	-0.396	0.116	0.190	0.253	0.370
15	-0.15	-0.582	-0.494	-0.445	-0.385	0.104	0.175	0.235	0.348
16	-0.15	-0.567	-0.482	-0.434	-0.376	0.093	0.161	0.219	0.329
17	-0.15	-0.554	-0.470	-0.424	-0.368	0.083	0.149	0.205	0.311
18	-0.15	-0.542	-0.460	-0.415	-0.360	0.075	0.139	0.193	0.296
19	-0.15	-0.531	-0.451	-0.407	-0.354	0.067	0.129	0.181	0.281
20	-0.15	-0.521	-0.442	-0.399	-0.348	0.060	0.120	0.171	0.268
21	-0.15	-0.512	-0.434	-0.393	-0.342	0.054	0.112	0.162	0.257
22	-0.15	-0.503	-0.427	-0.386	-0.337	0.048	0.105	0.153	0.246
23	-0.15	-0.495	-0.421	-0.380	-0.332	0.043	0.098	0.145	0.236
24	-0.15	-0.487	-0.414	-0.375	-0.328	0.038	0.092	0.138	0.226
11	-0.16	-0.667	-0.572	-0.517	-0.447	0.157	0.245	0.317	0.449
12	-0.16	-0.644	-0.551	-0.497	-0.431	0.137	0.219	0.288	0.415
13	-0.16	-0.623	-0.532	-0.481	-0.416	0.120	0.198	0.264	0.386
14	-0.16	-0.605	-0.516	-0.466	-0.404	0.106	0.180	0.243	0.361
15	-0.16	-0.589	-0.502	-0.454	-0.394	0.093	0.165	0.225	0.339
16	-0.16	-0.574	-0.490	-0.442	-0.385	0.083	0.151	0.209	0.319
17	-0.16	-0.561	-0.478	-0.433	-0.377	0.073	0.139	0.195	0.302
18	-0.16	-0.549	-0.468	-0.424	-0.369	0.065	0.129	0.183	0.286
19	-0.16	-0.538	-0.459	-0.415	-0.363	0.057	0.119	0.171	0.272
20	-0.16	-0.528	-0.450	-0.408	-0.357	0.050	0.110	0.161	0.259
21	-0.16	-0.519	-0.443	-0.401	-0.351	0.044	0.102	0.152	0.247
22	-0.16	-0.510	-0.436	-0.395	-0.346	0.038	0.095	0.143	0.236
23	-0.16	-0.503	-0.429	-0.389	-0.341	0.033	0.088	0.135	0.226
24	-0.16	-0.495	-0.423	-0.384	-0.337	0.028	0.082	0.128	0.217
11	-0.17	-0.673	-0.579	-0.525	-0.455	0.147	0.235	0.308	0.441
12	-0.17	-0.650	-0.558	-0.505	-0.439	0.127	0.210	0.279	0.406
13	-0.17	-0.629	-0.540	-0.489	-0.425	0.110	0.189	0.255	0.377
14	-0.17	-0.611	-0.524	-0.474	-0.413	0.096	0.171	0.234	0.352
15	-0.17	-0.595	-0.510	-0.462	-0.403	0.083	0.155	0.216	0.330
16	-0.17	-0.581	-0.497	-0.451	-0.393	0.072	0.141	0.200	0.310
18	-0.17	-0.556	-0.476	-0.432	-0.378	0.054	0.118	0.173	0.277
19	-0.17	-0.546	-0.467	-0.424	-0.371	0.047	0.109	0.161	0.262
20	-0.17	-0.536	-0.459	-0.417	-0.365	0.040	0.100	0.151	0.249
21	-0.17	-0.527	-0.451	-0.410	-0.360	0.034	0.092	0.142	0.237

Table 1, continued.

<i>N</i>	<i>Tau</i>	<i>L 99% CI</i>	<i>L 95% CI</i>	<i>L 90% CI</i>	<i>L 80% CI</i>	<i>U 80% CI</i>	<i>U 90% CI</i>	<i>U 95% CI</i>	<i>U 99% CI</i>
22	-0.17	-0.518	-0.444	-0.404	-0.355	0.028	0.084	0.133	0.226
23	-0.17	-0.510	-0.437	-0.398	-0.350	0.022	0.078	0.125	0.216
24	-0.17	-0.503	-0.431	-0.393	-0.346	0.018	0.071	0.118	0.207
11	-0.18	-0.679	-0.586	-0.532	-0.464	0.137	0.225	0.298	0.432
12	-0.18	-0.656	-0.565	-0.513	-0.447	0.117	0.200	0.269	0.398
13	-0.18	-0.635	-0.547	-0.496	-0.433	0.100	0.179	0.245	0.368
14	-0.18	-0.618	-0.531	-0.482	-0.422	0.085	0.160	0.224	0.343
15	-0.18	-0.602	-0.517	-0.470	-0.411	0.073	0.145	0.206	0.321
16	-0.18	-0.588	-0.505	-0.459	-0.402	0.062	0.131	0.190	0.301
17	-0.18	-0.575	-0.494	-0.449	-0.394	0.053	0.119	0.176	0.283
18	-0.18	-0.563	-0.484	-0.440	-0.387	0.044	0.108	0.163	0.267
19	-0.18	-0.553	-0.475	-0.432	-0.380	0.036	0.098	0.151	0.253
20	-0.18	-0.543	-0.467	-0.425	-0.374	0.030	0.090	0.141	0.240
21	-0.18	-0.534	-0.459	-0.418	-0.369	0.023	0.082	0.131	0.228
22	-0.18	-0.526	-0.452	-0.412	-0.364	0.017	0.074	0.123	0.217
23	-0.18	-0.518	-0.446	-0.407	-0.359	0.012	0.067	0.115	0.206
24	-0.18	-0.510	-0.440	-0.401	-0.355	0.007	0.061	0.107	0.197
11	-0.19	-0.684	-0.593	-0.539	-0.472	0.127	0.215	0.289	0.424
12	-0.19	-0.661	-0.572	-0.520	-0.455	0.106	0.190	0.260	0.389
13	-0.19	-0.642	-0.554	-0.504	-0.442	0.089	0.169	0.235	0.359
14	-0.19	-0.624	-0.539	-0.490	-0.430	0.075	0.150	0.214	0.334
15	-0.19	-0.609	-0.525	-0.478	-0.420	0.063	0.135	0.196	0.311
16	-0.19	-0.595	-0.513	-0.467	-0.411	0.052	0.121	0.180	0.291
17	-0.19	-0.582	-0.502	-0.457	-0.403	0.042	0.109	0.165	0.274
18	-0.19	-0.570	-0.492	-0.449	-0.396	0.034	0.098	0.153	0.258
19	-0.19	-0.560	-0.483	-0.441	-0.389	0.026	0.088	0.141	0.243
20	-0.19	-0.550	-0.475	-0.433	-0.383	0.019	0.079	0.131	0.230
21	-0.19	-0.541	-0.467	-0.427	-0.378	0.013	0.071	0.121	0.218
22	-0.19	-0.533	-0.460	-0.421	-0.373	0.007	0.064	0.113	0.207
23	-0.19	-0.525	-0.454	-0.415	-0.368	0.002	0.057	0.105	0.196
24	-0.19	-0.518	-0.448	-0.410	-0.364	-0.003	0.051	0.097	0.187
11	-0.20	-0.690	-0.600	-0.547	-0.480	0.117	0.205	0.279	0.415
12	-0.20	-0.667	-0.579	-0.528	-0.464	0.096	0.180	0.250	0.380
13	-0.20	-0.648	-0.561	-0.512	-0.450	0.079	0.158	0.225	0.350
14	-0.20	-0.630	-0.546	-0.498	-0.438	0.065	0.140	0.204	0.324
15	-0.20	-0.615	-0.532	-0.486	-0.428	0.052	0.124	0.186	0.302
16	-0.20	-0.601	-0.520	-0.475	-0.419	0.042	0.111	0.170	0.282
17	-0.20	-0.589	-0.510	-0.466	-0.411	0.032	0.099	0.155	0.264
18	-0.20	-0.577	-0.500	-0.457	-0.404	0.023	0.088	0.143	0.248
19	-0.20	-0.567	-0.491	-0.449	-0.398	0.016	0.078	0.131	0.233
20	-0.20	-0.557	-0.483	-0.442	-0.392	0.009	0.069	0.121	0.220
22	-0.20	-0.540	-0.468	-0.429	-0.382	-0.003	0.054	0.102	0.197
23	-0.20	-0.533	-0.462	-0.424	-0.377	-0.009	0.047	0.094	0.186
24	-0.20	-0.526	-0.456	-0.419	-0.373	-0.014	0.040	0.087	0.177

References

- American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.
- Bonett, D. G., & Wright, T. A. (2000). Sample size requirements for estimating Pearson, Kendall and Spearman correlations. *Psychometrika*, 65(1), 23-28. doi: [10.1007/BF02294183](https://doi.org/10.1007/BF02294183)
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(1), 1304-1312. doi: [10.1037/0003-066X.45.12.1304](https://doi.org/10.1037/0003-066X.45.12.1304)
- Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis*. New York, NY: Russell Sage Foundation.
- Fieller, E. C., Hartley, H. O., & Pearson, E. S. (1957). Tests for rank correlation coefficients. *Biometrika*, 44(3/4), 470-481. doi: [10.2307/2332878](https://doi.org/10.2307/2332878)
- Fisher, R. A. (1925). *Statistical methods for research workers*. London: Hafner Press.
- Gilpin, A. R. (1993). Table for conversion of Kendall's tau to Spearman's rho within the context of measures of magnitude of effect for meta-analysis. *Educational and Psychological Measurement*, 53(1), 87-92. doi: [10.1177/0013164493053001007](https://doi.org/10.1177/0013164493053001007)
- Helsel, D. R., & Hirsch, R. M. (1995). *Statistical methods in water resources*. Amsterdam: Elsevier Science.
- Kendall, M. G. (1949). Rank and product-moment correlation. *Biometrika*, 36(1/2), 177-193. doi: [10.2307/2332540](https://doi.org/10.2307/2332540)
- Levin, J. R., & Robinson, D. H. (2003). The trouble with interpreting statistically nonsignificant effect sizes in single-study investigations. *Journal of Modern Applied Statistical Methods*, 2(1), 231-236. Retrieved from <http://digitalcommons.wayne.edu/jmasm/vol2/iss1/23/>
- Long, J. D., & Cliff, N. (1997). Confidence intervals for Kendall's tau. *British Journal of Mathematical and Statistical Psychology*, 50(1), 31-41. doi: [10.1111/j.2044-8317.1997.tb01100.x](https://doi.org/10.1111/j.2044-8317.1997.tb01100.x)
- Rupinski, M. T., & Dunlap, W. P. (1996). Approximating Pearson product-moment correlations from Kendall's tau and Spearman's rho. *Educational and Psychological Measurement*, 56(3), 419-429. doi: [10.1177/0013164496056003004](https://doi.org/10.1177/0013164496056003004)

DAVID A. WALKER

Tukey, J. W. (1960). Conclusions vs decisions. *Technometrics*, 2(4), 423-433. doi: [10.1080/00401706.1960.10489909](https://doi.org/10.1080/00401706.1960.10489909)

Walker, D. A. (2015). *Confidence intervals for Kendall's tau with small samples*. [Computer program]. DeKalb, IL: Author.

JMASM Algorithms and Code Algorithm for Combining Robust and Bootstrap In Multiple Linear Model Regression (SAS)

Wan Muhamad Amir
University of Science, Malaysia
Kelantan, Malaysia

Mohamad Shafiq
University of Science, Malaysia
Kelantan, Malaysia

Hanafi A. Rahim
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Puspa Liza
Universiti Sultan Zainal Abidin
Kuala Terengganu, Malaysia

Azlida Aleng
University Malaysia Terengganu
Kuala Terengganu, Malaysia

Zailani Abdullah
University Malaysia Kelantan
Kelantan, Malaysia

The aim of bootstrapping is to approximate the sampling distribution of some estimator. An algorithm for combining method is given in SAS, along with applications and visualizations.

Keywords: Multiple linear regression, robust regression and bootstrap method

Introduction

Multiple linear regression (MLR) is an extension of simple linear regression. Table 1 displays the data for multiple linear regression.

Table 1. Data template for multiple linear regression

i	y_i	x_{i0}	x_{i1}	x_{i2}	...	x_{ip}
1	y_1	1	x_{11}	x_{12}	...	x_{1p}
2	y_2	1	x_{21}	x_{22}	...	x_{2p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
n	y_n	1	x_{n1}	x_{n2}	...	x_{np}

Dr. Amir bin W Ahmad is an Associate Professor of Biostatistics. Email him at: wmamir@usm.my. Mohamad Shafiq Bin Mohd Ibrahim is a postgraduate student in the School of Dental Sciences. Email him at: shafiqmat786@gmail.com.

MLR is used when there are two or more independent variables where the model using population information is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (1)$$

where β_0 is the intercept parameter and $\beta_0, \beta_1, \beta_2, \dots, \beta_{k-1}$ are the parameters associated with $k - 1$ predictor variables. The dependent variable $\mathbf{Y}$ is now written as a function of k independent variables, $x_1, x_2, \dots, x_k$.

The random error term is added to make the model probabilistic rather than deterministic. The value of the coefficient β_i determines the contribution of the independent variable x_i , and β_0 is the y -intercept. (Ngo, 2012). The coefficients $\beta_0, \beta_1, \dots, \beta_k$ are usually unknown because they represent population parameters. Below is the data presentation for multiple linear regression. General linear model in matrix form can be defined by the following vectors and matrices as below:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_{p-1} \end{bmatrix}$$

Calculation for Linear Regression using SAS

```
/* First we do simple linear regression */
proc reg data = temp1;
model y = x;
run;
```

Approach the MM-Estimation Procedure for Robust Regression

```
/* Then we do robust regression, in this case, MM-estimation */
proc robustreg data = temp1 method = MM;
model y = x;
run;
```

Procedure for Bootstrap with Case Resampling $n = 1000$

```
/* And finally we use a bootstrap with case resampling */
ods listing close;
proc surveyselect data = temp1 out = boot1 method = urs samprate =
1 outhits rep=1000;
run;
```


ALGORITHM FOR COMBINING IN MULTIPLE LINEAR REGRESSION

```
proc reg data = boot1 outest = est1(drop = _:);  
model y = x;  
by replicate; run;  
ods listing;
```

An Illustration of a Medical Case

Case Study I: A Case Study of Triglycerides

Table 2. Description of the variables

Variables	Code	Description
Triglycerides	Y	Triglycerides level of patients (mg/dl)
Weight	X1	Weight (kg)
Total Cholesterol	X2	Total cholesterol of patients (mg/dl)
Proconvertin	X3	Proconvertin (%)
Glucose	X4	Glucose level of patients (mg/dl)
HDL-Cholesterol	X5	High density lipoprotein cholesterol (mg/dl)
Hip	X6	Hip circumference (cm)
Insulin	X7	Insulin level of patients (IU/ml)
Lipid	X8	Taking lipid lowering medication (0 = no, 1 = yes)

Sources: Ahmad and Ibrahim (2013), Ahmad, Ibrahim, Halim, and Aleng (2014)

Algorithm for Combining Robust and Bootstrap in Multiple Linear Model Regression

Title 'Alternative Modeling on Multiple linear regression';

Data Medical;

Input Y X1 X2 X3 X4 X5 X6 X7 X8;

Datalines;

168	85.77	209	110	114	37	130.0	17	0
304	58.98	228	111	153	33	105.5	28	1
72	33.56	196	79	101	69	88.5	6	0
119	49.00	281	117	95	38	104.2	10	1
116	38.55	197	99	110	37	92.0	12	0
87	44.91	184	131	100	45	100.5	18	0
136	48.09	170	96	108	37	96.0	13	1
78	69.43	163	89	111	39	103.0	8	0
223	47.63	195	177	112	39	95.0	15	0
200	55.35	218	108	131	31	104.0	33	1
159	59.66	234	112	174	55	114.0	14	0
181	68.97	262	152	108	44	114.5	20	1
134	51.49	178	127	105	51	100.0	21	0
162	39.69	248	135	92	63	93.0	9	1

96	56.58	210	122	105	56	103.4	6	0
117	63.48	252	125	99	70	104.2	10	0
106	66.70	191	103	101	32	103.3	16	0
120	74.19	238	135	142	50	113.5	14	1
119	60.12	169	98	103	33	114.0	13	0
116	36.60	221	113	88	60	94.3	11	1
109	56.40	216	128	90	49	107.1	13	0
105	35.15	157	114	88	35	95.0	12	0
88	50.13	192	120	100	54	100.0	11	0
241	56.49	206	137	148	79	113.0	14	1
175	57.39	164	108	104	42	103.0	15	0
146	43.00	209	116	93	64	97.0	13	0
199	48.04	219	104	158	44	97.0	11	0
85	41.28	171	92	86	64	95.4	5	0
90	65.79	156	80	98	54	98.5	11	1
87	56.90	247	128	95	57	106.3	9	0
103	35.15	257	121	111	69	89.5	13	0
121	55.12	138	108	104	36	109.0	13	0
223	57.17	176	112	121	38	114.0	32	0
76	49.45	174	121	89	47	101.0	8	0
151	44.46	213	93	116	45	99.0	10	1
145	56.94	228	112	99	44	109.0	11	0
196	44.00	193	107	95	31	96.5	12	0
113	53.54	210	125	111	45	105.5	19	0
113	35.83	157	100	92	55	95.0	13	0

;

Run;

ods rtf file='results_ex1.rtf';

/* This first step is to make the selection of the data that have a significant impact with triglyceride levels. The next step is performing the procedure of modeling linear regression model */

```
proc reg data= Medical;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
run;
```

/* Then do robust regression, in this case MM-estimation */

```
proc robustreg data= Medical method=MM;
model Y = X1 X2 X3 X4 X5 X6 X7 X8/ diagnostics leverage;
```

ALGORITHM FOR COMBINING IN MULTIPLE LINEAR REGRESSION

```
output out=robout r=resid sr=stdres;
run;

/* Use a bootstrap with case resampling */

ods listing close;
proc surveyselect data= Medical out=boot1 method=urs samprate=1 outhits
rep = 50;
run;

/* And finally use a bootstrap with robust with case resampling */
proc robustreg data=boot1 method=MM plot=fitplot(nolimits) plots=all;
model Y = X1 X2 X3 X4 X5 X6 X7 X8;
run;

ods rtf close;
```

Results from Original Data

Below are the results from the analysis using the original data. The residual plots do not indicate any problem with the model. A normal distribution appears to fit our sample data fairly well. The plotted points form a reasonably straight line. In our case, the residual bounce randomly around the 0 line (residual vs. predicted value). This suggest that the assumption that the relationship is linear is reasonable. A higher R-squared value of 0.62 indicated how well the data fit the model and also indicates a better model.

Table 3. Parameter estimates for original data

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-86.5654	102.93662	-0.84	0.4070
x1	1	-1.08598	0.95288	-1.14	0.2634
x2	1	-0.06448	0.21973	-0.29	0.7712
x3	1	0.61857	0.36615	1.69	0.1015
x4	1	1.10882	0.33989	3.26	0.0028
x5	1	-0.52289	0.57119	-0.92	0.3673
x6	1	0.81327	1.38022	0.59	0.5601
x7	1	2.77339	1.25026	2.22	0.0343
x8	1	22.40585	14.51449	1.54	0.1331

Figure 1. Fit diagnostic for y

Figure 2. Outlier and Leverage Diagnostic for y

ALGORITHM FOR COMBINING IN MULTIPLE LINEAR REGRESSION

From Figure 2, we can see that there is no detection of outlier in observations. The leverage plots available in the SAS software are considered useful and effective in detecting multicollinearity, non-linearity, significance of the slope, and outliers (Lockwood & Mackinnon, 1998). Both of figures above indicate that this sample have no peculiarity and a data entry have no error. Figure 2 presented a regression diagnostics plot (a plot of the standardized residuals of robust regression MM versus the robust distance). Observations 2, 9, 10, 11, 18, 24, 27 and 33 are identified as leverage points. Below is the results of bootstrapping with $n = 50$:

Figure 3. Fit diagnostic for y after bootstrapping

Table 4 shows the results by using bootstrapping method. The aim of bootstrapping procedure is to approximate the entire sampling distribution of some estimator by resampling (simple random sampling with replacement) from the original data (Yaffee, 2002). The next step is to calculate the efficiency of the

bootstrap method with the original sample data. Table 5 summarize the findings of the calculated parameter.

Table 4. Parameter estimates using bootstrapping method

Parameter	DF	Estimate	Parameter Estimates		Chi-Square	Pr > ChiSq
			Standard Error	95% Confidence Limits		
Intercept	1	-297.0810	9.18120	-315.0760 -279.0860	1047.02	<0.0001
x1	1	-1.3526	0.07910	-1.5076 -1.1977	292.69	<0.0001
x2	1	0.0286	0.01850	-0.0077 0.0649	2.38	0.1227
x3	1	0.0441	0.04360	-0.0413 0.1295	1.03	0.3112
x4	1	1.5405	0.03300	1.4759 1.6052	2182.31	<0.0001
x5	1	0.2976	0.04960	0.2004 0.3948	36.04	<0.0001
x6	1	2.6234	0.12240	2.3836 2.8632	459.66	<0.0001
x7	1	2.4174	0.10580	2.2100 2.6248	521.88	<0.0001
x8	1	24.6443	1.20480	22.2829 27.0057	418.39	<0.0001
Scale	0	27.6976				

Table 5. Comparison of parameter estimates original sample and bootstrapping method

Variables	Original sample			Bootstrapping Method			Efficiency of Parameter (%)
	Parameter Estimate	Standard Error	P value	Estimate	Standard Error	P value	
Intercept	-86.56544	102.93662	0.4070	-297.0810	9.1812	< 0.0001	
x1	-1.08598	0.95288	0.2634	-1.3526	0.0791	< 0.0001	24.55
x2	-0.06448	0.21973	0.7712	0.0286	0.0185	0.1227	144.35
x3	0.61857	0.36615	0.1015	0.0441	0.0436	0.3112	92.87
x4	1.10882	0.33989	0.0028	1.5405	0.0330	< 0.0001	38.93
x5	-0.52289	0.57119	0.3673	0.2976	0.0496	< 0.0001	156.91
x6	0.81327	1.38022	0.5601	2.6234	0.1224	< 0.0001	222.57
x7	2.77339	1.25026	0.0343	2.4174	0.1058	< 0.0001	12.83
x8	22.40585	14.51449	0.1331	24.6443	1.2048	< 0.0001	9.99

References

Ahmad, W. M. A. W., & Ibrahim, M. S. (2013). High density lipoprotein cholesterol predicts triglycerides level in three distinct phases of blood pressure. *International Journal of Sciences: Basic and Applied Research*, 10(1), 38-46.
Retrieved from

ALGORITHM FOR COMBINING IN MULTIPLE LINEAR REGRESSION

<http://gssrr.org/index.php?journal=JournalOfBasicAndApplied&page=article&op=view&path%5B%5D=1111&path%5B%5D=1098>

Ahmad, W. M. A. W., Ibrahim, M. S., Halim, N., & Aleng, N. A. (2014). A study of triglycerides level in three distinct phases of human blood pressure: A case study from previous projects. *Applied Mathematical Sciences*, 8(46), 2289-2305. doi: 10.12988/ams.2014.42145

Lockwood, C. M., & Mackinnon, D. P. (1998). Bootstrapping the standard error off the mediated effect. *Proceedings of the 23rd Annual Meeting of SAS Users Group International*. Cary, NC: SAS Institute, Inc.

Ngo, T. H. D. (2012). The steps to follow in a multiple regression analysis. *Proceedings of the SAS Global Forum 2012 Conference (paper 333-2012)*. Cary, NC: SAS Institute Inc. Retrieved from <http://support.sas.com/resources/papers/proceedings12/333-2012.pdf>

Yaffee, R. A. (2002). *Robust regression analysis: Some popular statistical package options*. Statistics, Social Science, and Mapping Group, New York University, NY. Retrieved from http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/RobustRegAnalysis.pdf

JMASM Algorithms and Code

Determination of Optimal Tightened Normal Tightened Plan Using a Genetic Algorithm

Sampath Sundaram
University of Madras
Chennai, India

Deepa S. Parthasarathy
SDNB Vaishnav College for Women
Chennai, India

Designing a tightened normal tightened sampling plan requires sample sizes and acceptance number with switching criterion. An evolutionary algorithm, the genetic algorithm, is designed to identify optimal sample sizes and acceptance number of a tightened normal tightened sampling plan for a specified consumer's risk, producer's risk, and switching criterion. Optimal sample sizes and acceptance number are obtained by implementing the genetic algorithm. Tables are reported for various choices of switching criterion, consumer's quality level, and producer's quality level.

Keywords: tightened normal tightened sampling plan, average outgoing quality, switching criterion, genetic algorithm

---

Introduction

Companies aiming to remain competitive in order to retain a market share in a global economy need to maintain quality standards of highest order. The importance of consumer protection in sectors like the pharmaceutical industry has resulted in the popularity of $c = 0$ attribute sampling plans. It is to be observed that use of any positive acceptance number in a sampling plan results in passing lots which are likely to have defective units in them.

However, in safety and compliance testing, an acceptance number of zero is particularly desirable. In situations involving expensive testing procedures, practitioners often tend to use a single sampling plan with a sample of smaller size and acceptance number zero. But a sampling plan of this kind may result in the rejection of an entire lot based on the presence of even a single non-conforming unit. Apart from this, acceptance probabilities tend to decrease very rapidly for smaller values of p , namely, the fraction nonconforming in the lot.

Dr. Sundaram is a Professor of Statistics. Email him at: sampath1959@yahoo.com. Dr. Parthasarathy is a Teaching Faculty. Email her at: spdeepa_2007@yahoo.co.in.

TNT SAMPLING PLAN

This results in an Operating Characteristic (OC) curve with very poor shape. Even though these limitations can be overcome by using a single sampling plan with $c \geq 1$, a double sampling plan, or a multiple sampling plan, these sampling plans require larger sample sizes resulting in prohibitively expensive situations. Hence, to deal with such situations, Calvin (1977) devised a sampling scheme called Tightened Normal Tightened (TNT) sampling scheme.

Soundararajan and Vijayaraghavan (1992) studied TNT schemes with acceptance number $c > 0$ and compared its efficiency over single and double sampling plans. Suresh and Balamurali (1994) developed a Tightened Normal Tightened $TNT(n; 0, 1)$ scheme which has a switching rule between two sampling plans with fixed sample size and two minimum acceptance numbers, namely, $c = 0$ and $c = 1$. Suresh and Ramkumar (1996) studied the selection of single sampling plans indexed through Maximum Allowable Average Outgoing Quality (MAAOQ). Vijayaraghavan and Soundararajan (1996) developed procedures for the selection of $TNT(n; c_1, c_2)$ indexed by (AQL, LQL) and (AQL, AOQL) under the application of a Poisson model. Balamurali (2001) studied the selection of sampling schemes indexed by crossover point for compliance testing. Here, AQL, LQL and AOQL stand for Acceptable Quality Level, Limiting Quality Level and Average Outgoing Quality Level respectively.

Recently, the question of identifying sampling plans based on certain optimality criterion is receiving the attention of researchers. Because most of the times optimality criterion based on quantities like Average Sample Number assume complicated forms identifying optimal sampling plans is not a straightforward job. However, the availability of high speed computers and the evolution of soft computing tools have opened up a new direction in this regard. Sampath and Deepa (2012) developed a genetic algorithm for the determination of optimal sample sizes and acceptance number of double sampling plans under a crisp situation, and Sampath and Deepa (2013) designed a genetic algorithm for the same problem in situations involving both randomness and impreciseness. In this paper, it is proposed to identify optimal sample sizes and acceptance number of a tightened normal tightened plan using a genetic algorithm. Organization of the paper is as follows: A brief description on the tightened normal tightened scheme is given, followed by a description of the various stages involved in the implementation of the genetic algorithm. Finally, computational results are given in the final section.

Tightened Normal Tightened Scheme

The Tightened Normal Tightened (TNT) plan is a sampling plan appropriate for use in compliance sampling as well as in other areas of acceptance sampling. The conditions under which tightened normal tightened scheme can be applied are explained below.

- (i) Production is in a steady state so that results of past, present, and future lots are broadly indicative of a continuing process.
- (ii) Lots are submitted substantially in the order of their production.
- (iii) Inspection is by attributes, with quality defined as p , the fraction nonconforming.

A TNT scheme is specified by tightened sample size n_1 (large), normal sample size n_2 (small), criterion for switching to normal inspection t , and criterion for switching to tightened inspection s . Usually, s is smaller than t . It is carried out starting with tightened inspection.

1. Inspect using tightened inspection, with larger sample size n_1 and acceptance number $c = 0$.
2. Switch to normal inspection when t lots in a row are accepted under tightened inspection.
3. Inspect using normal inspection, with smaller sample size n_2 and acceptance number $c = 0$.
4. Switch to tightened inspection after a rejection if an additional lot is rejected in the next s lots.

The operating procedure for the above scheme, denoted by $TNT(n_1, n_2; 0)$, is based on the switching rule of United States Department of Defense (1963) with $s = 4$ and $t = 5$. One can refer to Dodge (1965), Hald and Thyregod (1965), and Stephens and Larson (1967) for derivation of composite OC function according to United States Department of Defense with the switching parameters $s = 4$ and $t = 5$. Let $P_1(p)$ be the probability of accepting a lot using tightened inspection and $P_2(p)$ be the probability of accepting a lot under normal inspection. The probability of accepting the lot is given by

$$P_a(p) = \frac{\delta P_1(p) + \mu P_2(p)}{\delta + \mu} \quad (1)$$

TNT SAMPLING PLAN

where

$$\delta(p) = \frac{1 - P_1^5(p)}{(1 - P_1(p)) \times P_1^5(p)} \quad (2)$$

is the average number of lots inspected on tightened inspection and

$$\mu(p) = \frac{2 - P_2^4(p)}{(1 - P_2(p)) \times (1 - P_2^4(p))} \quad (3)$$

is the average number of lots inspected on normal inspection.

When a Poisson model is implemented,

$$P_1(p) = \sum_{x=0}^c \frac{e^{-n_1 p} (n_1 p)^x}{x!}$$

and

$$P_2(p) = \sum_{x=0}^c \frac{e^{-n_2 p} (n_2 p)^x}{x!}$$

Calvin (1977) devised the OC function of the TNT scheme as

$$P_a(p) = \frac{P_1(1 - P_2^s)(1 - P_1^t)(1 - P_2) + P_2 P_1^t (1 - P_1)(2 - P_2^s)}{(1 - P_2^s)(1 - P_1^t)(1 - P_2) + P_1^t (1 - P_1)(2 - P_2^s)} \quad (4)$$

The composite OC curve, normal OC curve, and tightened OC curve of the TNT scheme $TNT(200, 100; 0)$ for $s = 4$ and $t = 5$ are as described in Figure 1.

Figure 1. Composite OC curve, normal OC curve, and tightened OC curve of the TNT scheme

A TNT plan is characterized by three parameters, namely, n_1 , n_2 , and c , with switching criterion s and t . One can determine the optimal parameters which satisfy the following two conditions for a specified producer's risk α , consumer's risk β , producer's quality level p_0 , consumer's quality level p_1 , s , and t .

$$\begin{aligned} P_{p_0}(\text{Accept Lot}) &\geq 1 - \alpha \\ P_{p_1}(\text{Accept Lot}) &\leq \beta \end{aligned} \quad (5)$$

or, equivalently,

$$\begin{aligned} P_{p_0}(\text{Accept Lot}) &\leq \alpha \\ P_{p_1}(\text{Accept Lot}) &\geq 1 - \beta \end{aligned} \quad (6)$$

It may be noted that there exists infinite number of solutions for n_1 , n_2 , and c satisfying (5) (or (6)). In order to obtain an optimal TNT plan, one has to define a suitable optimality criterion. In acceptance sampling, optimal sampling plans are determined based on measures of performance such as Average Sample Number,

TNT SAMPLING PLAN

Average Outgoing Quality (AOQ), and Average Total Inspection (ATI). In this paper, the problem of developing an optimal tightened normal tightened plan that minimizes the AOQ subject to the condition (5) (or (6)) is considered.

Average Outgoing Quality

In acceptance sampling programs, when the lots are rejected, they require some corrective actions in the form of replacement or elimination through 100 percent inspection. Such programs are known as rectifying inspection programs. AOQ is widely used for the evaluation of rectifying inspection, and represents average value of the lot quality that would be obtained over a long sequence of lots from a process with fraction defective p . AOQ for a TNT plan (Schilling and Neubauer, 2008) is given by

$$AOQ = p_0 P_a(p_0) \left(\frac{N - \bar{n}}{N} \right) \quad (7)$$

where

$$\bar{n} = \frac{n_1 (1 - P_2^s) (1 - P_1^t) (1 - P_2) + n_2 P_1^t (1 - P_1) (2 - P_2^s)}{(1 - P_2^s) (1 - P_1^t) (1 - P_2) + P_1^t (1 - P_1) (1 - P_2^s)}$$

and the lot size N is taken as $(n_1 + n_2) \times 10$ following Naidu et al. (2006).

Note that for a specified producer's risk α , consumer's risk β , producer's quality level p_0 , consumer's quality level p_1 , s , and t , the expressions for $P_a(p_0)$ and $P_a(p_1)$ are functions of n_1 , n_2 , and c . Hence solving for these sampling plan parameters such that (5) (or (6)) hold good becomes a complicated process. We therefore intend to make use of an unconventional algorithm like a genetic algorithm. The algorithm looks at a solution for n_1 , n_2 , and c such that (7) is minimum subject to the condition (5) (or (6)). The various steps associated with a genetic algorithm meant for solving the above problem are given in the following section.

Genetic Algorithm

Genetic algorithms (GAs) are evolutionary algorithms designed using the principle called Survival of the Fittest. These algorithms were first pioneered by Holland (1975). Genetic algorithms find their application in many fields, such as science,

engineering, business, and social sciences. Genetic algorithms are a domain independent problem solving approach and are very effective in identifying the optimal solution to a given problem. Details on the mechanism of GAs can be found in Goldberg (1989).

Genetic algorithms randomly search feasible points in a solution space in order to obtain best possible solution. It starts with the definition of what is known as population, which is made up of points representing different regions of the feasible solution space to the maximum extent possible. Each member in the given population is represented in the form of a string called a chromosome, and characters in a string are referred as genes. Defining a chromosome depends on the nature of the given problem. Fitness of a chromosome is determined by evaluating its objective function, namely the function being optimized, which indicates the nature of the solution as well as closeness towards optimality. A genetic algorithm tries to identify the best chromosome by successive breeding of existing chromosomes. Implementation of a genetic algorithm involves five different stages are explained below.

Defining initial population is the first stage of the genetic algorithm. Sets of chromosomes are formed in such a way that each chromosome produces one possible solution for the given optimization problem. Each chromosome defined in the initial population must be distinct in order for the GA to result in better solution. In this study, the initial population consists of 50 randomly generated chromosomes satisfying the probabilistic constraints given in (5) (or (6)). Each chromosome is comprised of nineteen genes. The first eight genes represent the binary encoding of the sample size n_1 , the second set of eight genes, i.e. from the ninth to the sixteenth bit position, represents the binary encoding of sample size n_2 , and the last three genes, the seventeenth, eighteenth, and nineteenth bit positions, gives the binary encoding of the parameter c . For example, if $n_1 = 130$, $n_2 = 100$, and $c = 2$, then the individual formation of the chromosome is as follows:

$$\overbrace{1\ 0\ 0\ 0\ 0\ 0\ 1\ 0}^{n_1} \overbrace{0\ 1\ 1\ 0\ 0\ 0\ 0\ 0}^{n_2} \overbrace{0\ 0\ 1\ 0}^c$$

Fitness value evaluation is the second stage of the genetic algorithm. For each chromosome existing in the initial population, the objective function corresponding to the given optimization problem is evaluated. These values are treated as fitness values. In this study, fitness values are computed by making use of the expression given in (7). Chromosomes having minimum AOQ value are treated as fitter.

TNT SAMPLING PLAN

Selection is the third stage of genetic algorithm. In this stage, chromosomes having high fitness value are selected to enter the mating pool with higher probabilities and a chromosome with lower fitness value is given a lower probability for entering the mating pool. Some of the selection procedures available in the literature are Roulette Wheel selection, Tournament selection, Ranking selection, and Proportional selection. In this paper, the Roulette Wheel selection procedure is used. For details related to selection procedures, one can refer to any standard text book on soft computing, such as Sivanandam and Deepa (2008).

Crossover is the fourth stage of genetic algorithm. In this stage, pair of chromosomes exist in the mating pool are combined to generate new chromosomes, called offspring. Many crossover mechanisms are available in the literature. In this work, a single point crossover mechanism is applied. In single point crossover, a crossover point is selected randomly in the interval $[1, l-1]$ where l is the length of a chromosome. The portions of the chromosome lying to the right of the crossover point are exchanged to produce offspring. For example, if

$$C_1: 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0$$

and

$$C_2: 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1$$

are two chromosomes with $l = 14$, the resulting offspring are

$$Ch_1: 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 1 \ 1 \ 1 \ 0 \ 0 \ 1$$

and

$$Ch_2: 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0,$$

respectively.

Mutation is the last stage of the algorithm. Each gene of a chromosome available in the given generation is randomly chosen and a small change is made with the help of mutation operator. In this work, each chromosome undergoes the following changes: A bit position is chosen randomly from the first eight bits (which is an encoding of sample size n_1) and its value is flipped. A second bit position is selected randomly from the ninth to sixteenth bit positions (which is an

encoding of sample size n_2) and its value is flipped. Similarly, flipping is done based on the last three bit positions (which is an encoding of acceptance number c). After mutation is over, mutated chromosomes form the new generation of chromosomes.

The five stages of the genetic algorithm described above are repeatedly performed. In this study, the number of times the above algorithm is to be repeated is taken as 50.

Determination of Optimal $TNT(n_1, n_2; c)$ Plan

The optimal TNT sampling plans for a wide range of p_0, p_1, s , and t with producer's risk $\alpha = 0.05$ and consumer's risk $\beta = 0.10$ are determined by implementing the genetic algorithm discussed earlier. The optimal sampling plans are displayed in Tables 1 to 4 in the appendix. The calculations are carried out using macros developed in Microsoft Excel VBA. The Microsoft Excel VBA codes developed in the determination of optimal TNT sampling plan are available from the authors.

Conclusion

A genetic algorithm has been designed and implemented for the determination of optimal $TNT(n_1, n_2; c)$ scheme. Various stages involved in a genetic algorithm are discussed in detail. Tables giving optimal sampling plans are constructed for various choices of s and t . The values are obtained using macros developed in Microsoft Excel VBA. It is observed that, for a specified $\alpha = 0.05$ and $\beta = 0.10$, acceptance number c increases when the producer's quality level p_0 increases. Also, the sample sizes n_1 and n_2 increase with increasing producer's quality level p_0 . It is to be noted that an increase in consumer's quality level p_1 decreases the sample sizes n_1 and n_2 . Also, the switching criterion s and t have no significant effect in minimum AOQ. That is, various choices of s and t considered in this study have almost the same effect in determining the optimal sampling plans.

References

- Balamurali, S. (2001). Selection of sampling schemes indexed by crossover point for compliance testing. *Economic Quality Control*, 16(1), 83-92. doi: [10.1515/EQC.2001.83](https://doi.org/10.1515/EQC.2001.83)

TNT SAMPLING PLAN

Calvin, T. W. (1977). TNT zero acceptance number sampling. In *American Society for Quality Control Annual Technical Conference Transactions* (35-39). Philadelphia, PA.

Dodge, H. F. (1965). *Evaluation of a sampling system having rules for switching between normal and tightened inspection: Technical report number 14*. Piscataway, NJ: Statistics Center, Rutgers University.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization and machine learning*. Reading, MA: Addison-Wesley Publishing Co.

Hald, A. & Thyregod, P. (1965). The composite operating characteristic under normal and tightened sampling inspection by attributes. *Bulletin of the International Statistical Institute*, 41, 517-529.

Holland, J. (1975). *Adaptation in natural and artificial system*. Ann Arbor, MI: University of Michigan Press.

Naidu, N. V. R., Babu, K. M., & Rajendra, G. (2006). *Total quality management*. New Delhi: New Age International.

Sampath, S. & Deepa, S. P. (2012). Determination of optimal double sampling plan using genetic algorithm. *Pakistan Journal of Statistics and Operations Research*, 8(2), 195-203. doi: [10.18187/pjsor.v8i2.255](https://doi.org/10.18187/pjsor.v8i2.255)

Sampath, S. & Deepa, S. P. (2013). Determination of optimal chance double sampling plan using genetic algorithm. *Model Assisted Statistics and Applications*, 8(4), 265-273. doi: [10.3233/MAS-130264](https://doi.org/10.3233/MAS-130264)

Schilling, E. G. & Neubauer, D. V. (2008). *Acceptance sampling in quality control* (2nd ed.). Boca Raton, FL: CRC Press.

Sivanandam, S. N. & Deepa, S. N. (2008). *Principles of soft computing*. New Delhi: Wiley India Private Limited.

Soundararajan, V. & Vijayaraghavan, R. (1992). Construction and selection of tightened-normal tightened sampling inspection scheme of type TNT-(n_1 , n_2 ; c). *Journal of Applied Statistics*, 19(3), 339- 349. doi: [10.1080/02664769200000031](https://doi.org/10.1080/02664769200000031)

Stephens, K. S. & Larson, K. E. (1967). An evaluation of the MIL-STD 105D system of sampling plans. *Industrial Quality Control*, 23(7), 310-319.

Suresh, K. K. & Balamurali, S. (1994). Construction and selection of tightened normal-tightened plans indexed by maximum allowable percent defective. *Journal of Applied Statistics*, 21(6), 589-595. doi: [10.1080/757584219](https://doi.org/10.1080/757584219)

Suresh, K. K. & Ramkumar, T. B. (1996). Selection of sampling plans indexed with maximum allowable average outgoing quality. *Journal of Applied Statistics*, 23(6), 645-654. doi: [10.1080/02664769623991](https://doi.org/10.1080/02664769623991)

United States Department of Defense. (1963). *Sampling procedures and tables for inspection by attributes* (MIL-STD 105D). Washington, DC: U.S. Government Printing Office.

Vijayaraghavan, R. & Soundararajan, V. (1996). Procedures and tables for the selection of tightened normal- tightened (TNT-(n; c1, c2)) sampling schemes. *Journal of Applied Statistics*, 23(1), 69-80. doi: [10.1080/02664769624369](https://doi.org/10.1080/02664769624369)

TNT SAMPLING PLAN

Appendix

Table 1. Optimal TNT plans for $s = 1$, $t = 2$ and 3 , $\alpha = 0.05$, and $\beta = 0.10$.

$p_0 = 0.001$					$s = 1, t = 2$				$s = 1, t = 3$			
					n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
p_1												
0.020					125	50	0	0.00092347	121	50	0	0.000922529
0.025					96	51	0	0.00091696	94	49	0	0.000919000
0.030					81	50	0	0.00091467	78	50	0	0.000913708
0.035					75	51	0	0.00091161	68	50	0	0.000910685
0.040					70	51	0	0.00091006	65	50	0	0.000909672
0.045					57	50	0	0.00090535	59	50	0	0.000907473
0.050					56	51	0	0.00090494	56	50	0	0.000906278

$p_0 = 0.002$					$s = 1, t = 2$				$s = 1, t = 3$			
					n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
p_1												
0.020					-	-	-	-	195	177	1	0.001810030
0.025					181	177	1	0.00180668	178	177	1	0.001805900
0.030					181	177	1	0.00180668	181	175	1	0.001809062
0.035					179	176	1	0.00180737	177	175	1	0.001808043
0.040					179	176	1	0.00180737	180	175	1	0.001808809
0.045					179	175	1	0.00180857	179	173	1	0.001810952
0.050					179	173	1	0.00181097	172	169	1	0.001813895

$p_0 = 0.003$					$s = 1, t = 2$				$s = 1, t = 3$			
					n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
p_1												
0.025					162	118	1	0.00273007	158	118	1	0.002727998
0.030					133	117	1	0.00271945	130	118	1	0.002715045
0.035					123	117	1	0.00271405	120	118	1	0.002709611
0.040					118	117	1	0.00271118	120	118	1	0.002709611
0.045					120	117	1	0.00271234	117	114	1	0.002718657
0.050					117	115	1	0.00271599	114	113	1	0.002719548

Table 1, continued.

$p_0 = 0.004$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	161	88	1	0.00366635	158	88	1	0.003663347	
0.030	200	199	2	0.00362244	198	195	2	0.003631958	
0.035	199	194	2	0.00363504	198	195	2	0.003631958	
0.040	199	197	2	0.00362723	197	194	2	0.003634081	
0.045	89	86	1	0.00362383	94	88	1	0.003619167	
0.050	89	88	1	0.00362142	89	86	1	0.003623784	
$p_0 = 0.005$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	-	-	-	-	158	69	1	0.004609512	
0.030	179	163	2	0.00452508	178	163	2	0.004524230	
0.035	171	163	2	0.00451987	166	163	2	0.004516433	
0.040	168	162	2	0.00452204	164	163	2	0.004515071	
0.045	162	160	2	0.00452618	163	161	2	0.004522712	
0.050	163	160	2	0.00452689	162	161	2	0.004522016	
$p_0 = 0.006$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.030	180	135	2	0.00546035	180	135	2	0.005459290	
0.035	162	136	2	0.00544026	153	135	2	0.005438235	
0.040	141	136	2	0.00542138	141	135	2	0.005427310	
0.045	137	135	2	0.00542346	135	134	2	0.005427414	
0.050	139	136	2	0.00541942	135	134	2	0.005427414	
$p_0 = 0.007$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.030	182	116	2	0.00639254	180	116	2	0.006388467	
0.035	198	195	3	0.00632088	199	195	3	0.006321612	
0.040	197	195	3	0.00632008	196	194	3	0.006325072	
0.045	122	116	2	0.00633116	195	192	3	0.006335817	
0.050	118	116	2	0.00632581	196	192	3	0.006336621	

TNT SAMPLING PLAN

Table 1, continued.

$p_0 = 0.008$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030	-	-	-	-	-	179	101	2	0.007325087
0.035	193	170	3	0.00724817		191	170	3	0.007245746
0.040	174	170	3	0.00722943		175	170	3	0.007230354
0.045	171	169	3	0.00723385		170	169	3	0.007232756
0.050	170	169	3	0.00723278		173	169	3	0.007235887

$p_0 = 0.009$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030	-	-	-	-	-	179	89	2	0.008268089
0.035	195	151	3	0.00818064		195	151	3	0.008178886
0.040	170	151	3	0.00815343		168	150	3	0.008160119
0.045	159	151	3	0.00813984		157	150	3	0.008146659
0.050	151	150	3	0.00813885		155	150	3	0.008144086

Table 2. Optimal TNT plans for $s = 1$, $t = 4$ and 5 , $\alpha = 0.05$, and $\beta = 0.10$.

$p_0 = 0.001$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
	0.020	118	50	0	0.000921715	116	50	0	0.000921029
	0.025	110	50	0	0.000920442	93	50	0	0.000916969
	0.030	82	49	0	0.000915993	87	50	0	0.000915651
	0.035	67	49	0	0.000911642	80	50	0	0.000913943
	0.040	58	49	0	0.000908415	70	50	0	0.000911125
	0.045	52	47	0	0.000908707	54	48	0	0.000908146
	0.050	49	45	0	0.000910168	49	47	0	0.000907333

Table 2, continued.

$p_0 = 0.002$					$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.020	196	177	1	0.001810177	195	177	1	0.001809868	195	177	1	0.001809868
0.025	176	175	1	0.001807781	185	177	1	0.001807581	185	177	1	0.001807581
0.030	176	175	1	0.001807781	184	177	1	0.001807344	184	177	1	0.001807344
0.035	176	175	1	0.001807781	179	177	1	0.001806137	179	177	1	0.001806137
0.040	175	173	1	0.001809911	175	174	1	0.001808714	175	174	1	0.001808714
0.045	175	172	1	0.001811102	175	173	1	0.001809903	175	173	1	0.001809903
0.050	173	172	1	0.001810583	177	172	1	0.001811595	177	172	1	0.001811595

$p_0 = 0.003$					$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025	158	117	1	0.002730294	158	117	1	0.002729820	158	117	1	0.002729820
0.030	134	117	1	0.002719641	133	117	1	0.002718982	133	117	1	0.002718982
0.035	119	117	1	0.002711724	126	117	1	0.002715464	126	117	1	0.002715464
0.040	116	114	1	0.002718050	116	115	1	0.002715363	116	115	1	0.002715363
0.045	115	112	1	0.002722804	115	114	1	0.002717454	115	114	1	0.002717454
0.050	113	111	1	0.002724269	114	113	1	0.002719532	114	113	1	0.002719532

$p_0 = 0.004$					$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025	162	87	1	0.003668336	157	87	1	0.003664266	157	87	1	0.003664266
0.030	200	198	2	0.003625040	200	198	2	0.003625024	200	198	2	0.003625024
0.035	200	198	2	0.003625040	200	198	2	0.003625024	200	198	2	0.003625024
0.040	198	195	2	0.003631937	198	195	2	0.003631915	198	195	2	0.003631915
0.045	198	192	2	0.003639654	198	195	2	0.003631915	198	195	2	0.003631915
0.050	199	197	2	0.003627200	196	194	2	0.003633585	196	194	2	0.003633585

$p_0 = 0.005$					$s = 1, t = 2$				$s = 1, t = 3$			
p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025	156	69	1	0.004604499	156	68	1	0.004607669	156	68	1	0.004607669
0.030	182	163	2	0.004526400	180	163	2	0.004524937	180	163	2	0.004524937
0.035	166	163	2	0.004516391	168	162	2	0.004521788	168	162	2	0.004521788
0.040	164	163	2	0.004515057	163	160	2	0.004526771	163	160	2	0.004526771
0.045	161	160	2	0.004525437	161	159	2	0.004529517	161	159	2	0.004529517
0.050	161	160	2	0.004525437	167	159	2	0.004533512	167	159	2	0.004533512

TNT SAMPLING PLAN

Table 2, continued.

$p_0 = 0.006$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025		156	56	1	0.005554394	156	56	1	0.005543875
0.030		183	135	2	0.005460090	178	135	2	0.005455578
0.035		158	135	2	0.005441970	153	135	2	0.005437430
0.040		136	135	2	0.005422414	135	134	2	0.005427375
0.045		136	135	2	0.005422414	134	133	2	0.005432312
0.050		136	135	2	0.005422414	136	130	2	0.005451835

$p_0 = 0.007$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030		181	115	2	0.006394767	185	115	2	0.006393767
0.035		198	194	3	0.006326581	198	195	3	0.006320708
0.040		196	194	3	0.006325035	196	194	3	0.006324995
0.045		194	193	3	0.006329237	196	193	3	0.006330752
0.050		193	192	3	0.006334178	196	193	3	0.006330752

$p_0 = 0.008$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030		183	100	2	0.007333215	180	100	2	0.007324785
0.035		196	170	3	0.007249456	193	170	3	0.007246189
0.040		174	170	3	0.007229235	182	170	3	0.007236643
0.045		172	170	3	0.007227226	177	170	3	0.007232004
0.050		173	169	3	0.007235791	171	169	3	0.007233710

$p_0 = 0.009$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030		178	88	2	0.008272107	179	87	2	0.008275323
0.035		193	151	3	0.008175165	192	150	3	0.008182135
0.040		170	151	3	0.008152119	168	150	3	0.008158866
0.045		162	151	3	0.008142936	150	149	3	0.008146914
0.050		160	151	3	0.008140541	153	149	3	0.008150742

Table 3. Optimal TNT plans for $s = 2$, $t = 3$ and 4, $\alpha = 0.05$, and $\beta = 0.10$.

$p_0 = 0.001$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.020	117	50	0	0.000921194	117	49	0	0.000921922	
0.025	100	50	0	0.000918348	98	49	0	0.000918885	
0.030	78	47	0	0.000917536	82	49	0	0.000915505	
0.035	71	47	0	0.000915554	66	48	0	0.000912430	
0.040	59	47	0	0.000911510	62	48	0	0.000911094	
0.045	56	47	0	0.000910345	55	48	0	0.000908488	
0.050	54	45	0	0.000912349	48	46	0	0.000908244	

$p_0 = 0.002$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.020	195	177	1	0.001809850	195	177	1	0.001809705	
0.025	181	177	1	0.001806620	178	176	1	0.001807066	
0.030	177	175	1	0.001808025	177	175	1	0.001808010	
0.035	176	174	1	0.001808964	177	174	1	0.001809192	
0.040	177	174	1	0.001809214	177	173	1	0.001810373	
0.045	177	174	1	0.001809214	174	172	1	0.001810819	
0.050	177	174	1	0.001809214	177	172	1	0.001811595	

$p_0 = 0.003$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	163	117	1	0.002731659	159	117	1	0.002729329	
0.030	131	117	1	0.002717968	134	117	1	0.002719125	
0.035	126	117	1	0.002715442	122	116	1	0.002715917	
0.040	118	116	1	0.002713820	119	116	1	0.002714328	
0.045	116	114	1	0.002718027	116	115	1	0.002715345	
0.050	115	114	1	0.002717451	115	114	1	0.002717435	

$p_0 = 0.004$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	156	87	1	0.003664103	157	86	1	0.003666564	
0.030	200	198	2	0.003625019	199	198	2	0.003624566	
0.035	198	197	2	0.003626737	198	195	2	0.003631868	
0.040	198	197	2	0.003626737	197	196	2	0.003628867	
0.045	199	197	2	0.003627179	197	195	2	0.003631434	
0.050	198	196	2	0.003629326	198	196	2	0.003629299	

TNT SAMPLING PLAN

Table 3, continued.

$p_0 = 0.005$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025		158	69	1	0.004602243	156	68	1	0.004602324
0.030		178	163	2	0.004523715	186	163	2	0.004527613
0.035		169	163	2	0.004518243	178	163	2	0.004523300
0.040		167	163	2	0.004516975	178	163	2	0.004523300
0.045		162	161	2	0.004521985	172	162	2	0.004524010
0.050		160	159	2	0.004528829	168	162	2	0.004521611

$p_0 = 0.006$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.025		-	-	-	-	160	55	1	0.005547309
0.030		183	135	2	0.005458675	178	135	2	0.005453510
0.035		164	135	2	0.005445872	154	135	2	0.005437400
0.040		137	135	2	0.005423333	139	135	2	0.005425051
0.045		136	134	2	0.005428327	138	134	2	0.005430076
0.050		138	133	2	0.005436135	135	134	2	0.005427335

$p_0 = 0.007$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030		178	115	2	0.006390131	179	114	2	0.006394389
0.035		199	195	3	0.006321429	198	195	3	0.006320584
0.040		196	194	3	0.006324985	196	195	3	0.006319167
0.045		194	193	3	0.006329213	195	193	3	0.006329910
0.050		194	193	3	0.006329213	197	192	3	0.006337049

$p_0 = 0.008$		$s = 1, t = 2$				$s = 1, t = 3$			
p_1		n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030		179	100	2	0.007326484	178	99	2	0.007327863
0.035		191	170	3	0.007244411	191	170	3	0.007243322
0.040		175	170	3	0.007230061	171	170	3	0.007226132
0.045		174	170	3	0.007229101	171	170	3	0.007226132
0.050		171	169	3	0.007233695	170	169	3	0.007232657

Table 3, continued.

p_1	$p_0 = 0.009$				$s = 1, t = 2$				$s = 1, t = 3$			
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.030	183	88	2	0.008268077	180	87	2	0.008264219	180	87	2	0.008264219
0.035	193	151	3	0.008173139	195	149	3	0.008190940	195	149	3	0.008190940
0.040	180	151	3	0.008161396	171	149	3	0.008170188	171	149	3	0.008170188
0.045	157	150	3	0.008146156	150	149	3	0.008146852	150	149	3	0.008146852
0.050	153	149	3	0.008150708	150	149	3	0.008146852	150	149	3	0.008146852

Table 4. Optimal TNT plans for $s = 2, t = 4$ and $5, \alpha = 0.05$, and $\beta = 0.10$.

p_1	$p_0 = 0.001$				$s = 1, t = 2$				$s = 1, t = 3$			
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.020	118	49	0	0.000921353	125	49	0	0.000922006	125	49	0	0.000922006
0.025	113	49	0	0.000920707	100	49	0	0.000918607	100	49	0	0.000918607
0.030	77	48	0	0.000915360	81	48	0	0.000916232	81	48	0	0.000916232
0.035	66	47	0	0.000913652	70	48	0	0.000913391	70	48	0	0.000913391
0.040	59	47	0	0.000911304	68	48	0	0.000912808	68	48	0	0.000912808
0.045	54	47	0	0.000909407	54	47	0	0.000909380	54	47	0	0.000909380
0.050	50	47	0	0.000907736	52	47	0	0.000908570	52	47	0	0.000908570

p_1	$p_0 = 0.002$				$s = 1, t = 2$				$s = 1, t = 3$			
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
0.020	195	177	1	0.001809553	195	177	1	0.001809493	195	177	1	0.001809493
0.025	182	177	1	0.001806781	187	177	1	0.001807844	187	177	1	0.001807844
0.030	179	176	1	0.001807280	186	177	1	0.001807631	186	177	1	0.001807631
0.035	178	176	1	0.001807051	180	176	1	0.001807495	180	176	1	0.001807495
0.040	177	174	1	0.001809169	177	175	1	0.001807988	177	175	1	0.001807988
0.045	175	173	1	0.001809871	176	173	1	0.001810098	176	173	1	0.001810098
0.050	174	173	1	0.001809633	173	172	1	0.001810560	173	172	1	0.001810560

TNT SAMPLING PLAN

Table 4, continued.

$p_0 = 0.003$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	157	117	1	0.002727763	157	117	1	0.002727495	
0.030	133	117	1	0.002718362	140	117	1	0.002721202	
0.035	120	117	1	0.002712145	123	117	1	0.002713613	
0.040	116	115	1	0.002715328	118	117	1	0.002711102	
0.045	116	115	1	0.002715328	116	115	1	0.002715321	
0.050	115	114	1	0.002717420	115	114	1	0.002717412	

$p_0 = 0.004$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	157	86	1	0.003663245	157	86	1	0.003662696	
0.030	200	198	2	0.003624961	198	197	2	0.003626702	
0.035	198	197	2	0.003626709	198	197	2	0.003626702	
0.040	198	196	2	0.003629271	195	194	2	0.003633095	
0.045	200	196	2	0.003630100	194	192	2	0.003637721	
0.050	197	195	2	0.003631407	198	196	2	0.003629259	

$p_0 = 0.005$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	157	66	1	0.004610771	156	67	1	0.004602109	
0.030	183	162	2	0.004529595	178	163	2	0.004522695	
0.035	163	162	2	0.004518474	165	163	2	0.004515560	
0.040	162	161	2	0.004521936	165	163	2	0.004515560	
0.045	162	161	2	0.004221936	164	163	2	0.004514975	
0.050	162	159	2	0.004530044	163	160	2	0.004526583	

$p_0 = 0.006$									
p_1	$s = 1, t = 2$				$s = 1, t = 3$				AOQ
	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ	
0.025	166	53	1	0.005550880	157	54	1	0.005544301	
0.030	183	134	2	0.005459981	185	134	2	0.005460246	
0.035	153	134	2	0.005441853	153	134	2	0.005441560	
0.040	136	134	2	0.005428185	143	134	2	0.005433987	
0.045	135	134	2	0.005427299	135	134	2	0.005427283	
0.050	132	131	2	0.005441986	135	134	2	0.005427283	

Table 4, continued.

$p_0 = 0.007$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
	0.030	179	114	2	0.006388881	180	114	2	0.006388371
	0.035	196	195	3	0.006319130	200	195	3	0.006321711
	0.040	196	195	3	0.006319130	199	195	3	0.006321071
	0.045	195	194	3	0.006324159	196	195	3	0.006319114
	0.050	195	194	3	0.006324159	196	194	3	0.006324813
$p_0 = 0.008$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
	0.030	179	98	2	0.007328779	178	98	2	0.007328104
	0.035	193	170	3	0.007243578	193	170	3	0.007243124
	0.040	177	170	3	0.007231282	175	169	3	0.007236929
	0.045	174	170	3	0.007228724	173	169	3	0.007235221
	0.050	172	170	3	0.007226973	173	169	3	0.007235221
$p_0 = 0.009$	$s = 1, t = 2$				$s = 1, t = 3$				
	p_1	n_1	n_2	c	AOQ	n_1	n_2	c	AOQ
	0.030	186	85	2	0.008271138	188	85	2	0.008272709
	0.035	195	150	3	0.008177279	197	150	3	0.008177442
	0.040	172	150	3	0.008160109	171	150	3	0.008158701
	0.045	152	150	3	0.008139789	156	149	3	0.008153466
	0.050	150	149	3	0.008146798	155	148	3	0.008161672

Statistical Software Applications & Review Analyzing Different Sampling Designs (SAS)

Ying Lu

Educational Testing Service
Rosedale Road, NJ

Various sampling designs are reviewed within the framework of probability sampling. SAS® code to estimate means and proportions, and their standard errors, using different sampling designs are illustrated using example data sets.

Keywords: Sampling, SAS

Introduction

Researchers and statisticians often find it necessary to apply survey-sampling methodologies to acquire information about a large population. Sampling can take different forms (e.g., simple random sampling, stratified sampling, or clustering sampling) and different levels. In order to make appropriate and statistically valid inferences about the population based on the selected sample, the sampling design needs to be taken into consideration in the data analysis.

The purpose of this article is to provide step-by-step guidance on how data obtained from various sampling designs could be using SAS (ver. 9.2) PROC SURVEYMEANS, and to illustrate how to estimate means and proportions, and their standard errors, in various finite sampling designs within the framework of probability sampling. In situations when it is less straightforward to use PROC SURVEYMEANS to obtain the estimates, the use of PROC IML as an alternative tool is illustrated. SAS/IML is an interface that provides interactive matrix programming. It is a separate component from SAS that may require additional installation.

For each sampling design listed below, a brief summary of the model and procedure including formulas for the estimated statistics, their variances, and approximate confidence intervals will be presented. The sampling designs

---

Ying Lu is a Senior Psychometrician. Email her at: ylu@ets.org.

discussed in this paper all belong to the category of sampling without replacement. Sampling-with-replacement designs are comparatively easy to analyze and therefore are not discussed here. More details and references for the sampling models, designs, and proofs for the formulas used in this paper, as well as definitions and terms adopted in this paper may be found in Lohr (1999). SAS features are demonstrated using example data sets, SAS programs, and output. Data sets used in the examples are selected from the CD accompanying the Lohr book so that interested readers can have access to them. This paper assumes that nonsampling errors such as selection bias and inaccuracy of responses can be ignored in the sampling designs. The sampling designs discussed in this paper are as follows:

- simple random sampling,
- stratified sampling with a Simple Random Sample (SRS) selected from each stratum,
- one-stage cluster sampling with an SRS of clusters,
- two-stage cluster sampling with an SRS at each stage,
- stratified sampling with one-stage cluster sampling (using SRS) within each stratum,
- one-stage cluster sampling with unequal probabilities,
- general complex surveys.

The notations used in this paper differ by section and sampling design. In the more general setting, consider U to be a finite population, and S to be a selected sample. Within the complex sampling framework, subscripts are added to S to denote the sample within a specific cluster or stratum.

Methodology

Simple Random Sampling

Estimating population mean and total Let y_i be the value of interest associated with the i^{th} unit in the population, and let $\bar{y}$ and s^2 denote the sample mean and sample variance, respectively. The population mean $\bar{y}_U$ in an SRS is estimated by the sample mean

$$\hat{\bar{y}}_U = \bar{y} = \frac{1}{n} \sum_{i \in S} y_i \quad (1)$$

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

with variance reported as

$$\hat{V}(\hat{\bar{y}}) = \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \quad (2)$$

and the population total is estimated by

$$\hat{t} = N\bar{y} \quad (3)$$

with variance estimated by

$$\hat{V}(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s^2}{n} . \quad (4)$$

For estimating the mean, PROC SURVEYMEANS can be run using only the option of 'total = N', which specifies the population size to enable a finite population correction. There are two ways of estimating the total. A weight variable of N/n can be created, and the sum option can be used. Notice that a common weight on all variables will not affect the estimation of the mean. Alternatively, a new variable can be created that equals $N*y$, where y is the original variable and the mean can be computed on this variable.

As an example, 'counties.dat' (Lohr, 1999, p. 440) is used to illustrate the estimation of the population mean and total. The data set contains information on land area, population, numbers of physicians, unemployment, and a number of other quantities for an SRS of 100 counties from the 3,141 counties in the United States. The mean and number of physicians are estimated, along with their associated standard errors and 95% confidence intervals. The SAS code used to obtain the estimates is provided below together with the output.

SAS code:

```
data counties;
infile 'C:\Sampling\counties.dat' dlm=',' firstobs=2;
input RN STATE $ COUNTY $ LANDAREA TOTPOP PHYSICIA ENROLL PERCPUB
CIVLABOR UNEMP FARMPOP NUMFARM FARMACRE FEDGRANT FEDCIV MILIT VETERANS
PERCVIET;
```

```

/* convert missing values */
array numval[*] _numeric_;
do i = 1 to dim(numval);
  if numval[i]=-99 then numval[i]=.;
end;
drop i;
run;

data srs;
set counties;
wt=3141/100;
run;

proc surveymeans data=srs mean clm sum clsum total=3141;
var PHYSICIA;
weight wt;
run;

```

The first part of the SAS code reads in the data and converts the missing value coding of ‘-99’ to the SAS default coding of a period. The second part of the code defines the weight variable as N/n . The SURVEYMEANS procedure is in the third part of the code with the *clm* and *clsum* options added to give confidence intervals as well as estimates for the mean and the total. The output of the program is displayed below:

The SURVEYMEANS Procedure

Data Summary

Number of Observations	100
Sum of Weights	3141

Statistics

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Sum	Std Dev
PHYSICIA	297.170000	156.632533	-13.622928	607.962928	933411	491983

Statistics

Variable	Lower 95% CL for Sum	Upper 95% CL for Sum
PHYSICIA	-42790	1909612

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

The SURVEYMEANS procedure produces estimated mean, the standard error (SE) of mean, and confidence intervals. The estimated population mean is $\hat{\bar{y}}_U = 297.17$ with $SE(\hat{\bar{y}}) = 156.6325$. The estimated population total is $\hat{t} = 933,411$ with $SE(\hat{t}) = 491,983$.

Estimating the proportion

As a special case of mean estimation, the population proportion p in an SRS is estimated by the sample proportion

$$\hat{p} = \bar{y} \quad (5)$$

and

$$SE(\hat{p}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{\hat{p}(1 - \hat{p})}{n - 1}} \quad (6)$$

For the estimation of the proportion, a variable needs to be created so that it takes the value of 1 if the unit has the characteristic of interest and 0 otherwise, and then PROC SURVEYMEANS can be run on the new variable. In the example below, we are interested in the percentage of children that are overdue for a vaccination in a school. Suppose the population consists of 120 children, and the selected sample consists of 10 children with 4 of them being overdue.

```
data a;
  input patient status $ @@;
  cards;
  1 ok 2 ok 3 ok 4 overdue 5 overdue
  6 ok 7 overdue 8 ok 9 overdue 10 ok
  ;;
run;

data a;
  set a;
  if status='overdue' then y=1;
  else y=0;
  wt=120/10;
run;

proc surveymeans data=a mean clm total=120;
var y;
weight wt;
```

run;

The SURVEYMEANS Procedure

Data Summary

Number of Observations	10
Sum of Weights	120

Statistics

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
y	0.400000	0.156347	0.046318	0.753682

Therefore the estimated proportion is 0.4000 with the SE being 0.1563.

Ratio Estimation

The use of ratio estimation requires measures of y_i and x_i on each sampling unit. The ratio of the two quantities is defined as

$$B = \frac{\bar{y}_U}{\bar{x}_U} . \quad (7)$$

It is estimated by

$$\hat{B} = \frac{\bar{y}}{\bar{x}} = \frac{\hat{t}_y}{\hat{t}_x} \quad (8)$$

with estimated variance

$$\begin{aligned}\hat{V}[\hat{B}] &= \left(1 - \frac{n}{N}\right) \frac{1}{n\bar{x}_U^2} \frac{\sum_{i \in S} (y_i - \hat{B}x_i)^2}{n-1} \\ &= \left(1 - \frac{n}{N}\right) \frac{s_e^2}{n\bar{x}_U^2}\end{aligned}\quad (9)$$

where $e_i = y_i - \hat{B}x_i$. When $\bar{x}_U$ is not known, we use $\bar{x}$, the sample mean, to approximate it.

Ratio estimation may not be run directly using PROC SURVEYMEANS. Two procedures are demonstrated here, using PROC SURVEYMEANS after some preliminary analyses that feed into it, or via the PROC IML procedure.

For PROC SURVEYMEANS to obtain the correct standard error for ratio estimation, we need to first create a new variable $d = \frac{y - \hat{B}x}{\bar{x}_U}$, and run PROC

SURVEYMEANS on d as in SRS. As an example, we use ‘counties.dat’ to estimate the average farm population per square mile of land area. The y variable is the farm population and the x variable is the land area. Here, $\bar{x}_U$ is assumed to be unknown, and therefore is approximated by $\bar{x}$. The SAS program applied is as follows:

```
proc means mean data=counties noway;
var FARMPOP LANDAREA;
run;

data ratio;
set counties;
d=(FARMPOP-LANDAREA*1.2137218)/944.92;
/* 944.92 is the sample mean of x obtained from proc means.
   1.213718 is the ratio estimate computed from proc means
   output. */
run;

proc surveymeans data=ratio total=3141;
var d;
run;
```

The first part of the program uses the MEANS procedure to get the sample means of x and y . The ratio estimate would be computed as $\bar{y}/\bar{x}$. Next the program creates the variable d , which is computed from the output we obtained from running the first part of the code. And lastly PROC SURVEYMEANS for SRS is run on the

variable d . The resulting SE for the mean of d , 0.1891, is the desired SE for the ratio estimate.

The MEANS Procedure

Variable	Mean
FARMPPOP	1146.87
LANDAREA	944.9200000

The SURVEYMEANS Procedure

Data Summary

Number of Observations	100

Statistics

Variable	N	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean
d	100	-3.445794E-9	0.189105	-0.375225	0.375225

Alternatively, the IML procedure in SAS can be used to calculate the ratio estimate and its SE according to the formulas given. The following SAS program accomplishes the same task as the previous one. The program reads in the number of physicians and population for each county into vectors $\mathbf{y}$ and $\mathbf{x}$, creates a new vector, $\mathbf{e}$, defined as

$$\mathbf{e} = \mathbf{y} - \frac{\bar{y}}{\bar{x}} \mathbf{x}$$

gets the variance of $\mathbf{e}$, and calculates the ratio estimate and SE according to the formula.

```
proc iml;
  use counties;
  /* define x and y */
  read all var{FARMPPOP} into y;
  read all var{LANDAREA} into x;
  close counties;
```

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

```
bign=3141; /* designates the population size N (bign) */
n=nrow(x); /* get sample size */
xbar=sum(x)/n;
ybar=sum(y)/n;
bhat=ybar/xbar; /* the ratio estimate*/
e=y-bhat*x;
vard=(ssq(e)-(sum(e)**2/n)/(n-1));
varbhat=((bign-n)/(bign*n))*vard/(xbar**2); /* estimated variance
for the estimate*/
sebhat=sqrt(varbhat); /* standard error */
print " Ratio Estimation";
print ybar xbar bhat varbhat sebhat;
quit;
run;
```

The output below gives a ratio estimate $\hat{B}$ of 1.2137 with the SE being 0.1891, which matches the results we obtained earlier using the first approach.

Ratio Estimation

YBAR	XBAR	BHAT	VARBHAT	SEBHAT
1146.87	944.92	1.2137218	0.0357606	0.1891048

Ratio estimation is also used to estimate the total and mean of a single variable to increase the precision of the estimates. Ratio estimation gives better performance than the regular estimation of the mean of y when y and the auxiliary variable x are linearly related, and specifically, when the data are well fit by a straight line through the origin. In the discussion of one-stage cluster sampling with an SRS of clusters later in the paper, an example is provided where the ratio estimate of the mean gives a smaller error variance than the unbiased estimate of the mean.

Regression Estimation

Although ratio estimation works best for data that are well fit by a straight line through the origin, regression estimation might be more suitable for data that scatter around a straight line with an intercept, which is modeled by $y = \beta_0 + \beta_1 x$.

With $\bar{x}_U$ assumed to be known, the regression estimator of $\bar{y}_U$ is determined by

$$\hat{y}_{\text{reg}} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_U \quad (10)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares regression coefficients.

The regression estimator is biased. Its standard error can be approximated as

$$\text{SE}(\hat{y}_{\text{reg}}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_e^2}{n}} \quad (11)$$

where $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$.

One way to obtain the standard error is to calculate s_e^2 from the residual sum of squares in the regression analysis output. Another possibility is to, as in ratio estimation, create the variable e_i , and run PROC SURVEYMEANS on the new variable. Lastly, the estimation can be implemented through the IML environment. Using ‘counties.dat’, treat population as the auxiliary variable and estimate the total number of physicians in the United States, along with the standard error. The 1993 United States total population was estimated to be 255,077,536. The regression procedure in SAS (PROC REG) is used to obtain the ANOVA table and regression coefficients:

```
proc reg data=counties;
  model PHYSICIA=TOTPOP;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: PHYSICIA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	239521351	239521351	2068.16	<.0001
Error	98	11349761	115814		
Corrected Total	99	250871112			

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

Root MSE	340.31440	R-Square	0.9548
Dependent Mean	297.17000	Adj R-Sq	0.9543
Coeff Var	114.51842		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-54.23128	34.89764	-1.55	0.1234
TOTPOP	1	0.00296	0.00006519	45.48	<.0001

with $\hat{t}_x = 255,077,536$, $\hat{\beta}_1 = 0.00296$, and $\hat{\beta}_0 = -54.23128$, the total number of physicians is estimated by

$$\hat{t}_{\text{yreg}} = \hat{t}_x \hat{\beta}_1 + N \hat{\beta}_0 = 255,077,536 \times 0.00296 \times 3141 \times (-54.23128) = 584689.0561$$

$$s_e = \sqrt{\frac{11349761}{n-1}} = \sqrt{\frac{11349761}{99}} = 338.5913 ,$$

so

$$\text{SE}(\hat{t}_{\text{yreg}}) = N \sqrt{1 - \frac{n}{N}} \frac{s_e}{\sqrt{n}} = 3141 \sqrt{1 - \frac{100}{3141}} \frac{338.5913}{\sqrt{100}} = 104644.8691 .$$

The same result can be obtained by creating $e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ and applying the regular SURVEYMEANS procedure to obtain the sum of e_i and associated error. Suppose PROC REG was previously conducted, and coefficients $\hat{\beta}_0$ (-54.2313) and $\hat{\beta}_1$ (0.0030) were obtained. The SAS code for obtaining the standard error of the regression estimator and its output are shown below:

```
data reg;
set counties;
wt=3141/100;
```

```

e=PHYSICIA-(-54.23128+TOTPOP*0.00296);
run;

proc surveymeans total=3141 sum;
var e;
weight wt;
run;

```

The SURVEYMEANS Procedure

Data Summary

Number of Observations	100
Sum of Weights	3141

Statistics

Variable	Sum	Std Dev
e	1724.554742	104648

Stratified Sampling with an SRS in Each Stratum

Stratified sampling means the population is divided into a number of mutually exclusive and exhaustive strata, and a probability sample is drawn from each stratum independently. The simplest form of stratified sampling where an SRS is taken from each stratum is of interest.

Let H denote the number of strata; N_h denote the number of sampling units in stratum h ; $\bar{y}_h$ denote the sample mean in stratum h ; and s_h^2 denote the sample variance in stratum h . The population total and mean are estimated by

$$\hat{t}_{\text{str}} = \sum_{h=1}^H N_h \bar{y}_h, \quad (12)$$

$$\bar{y}_{\text{str}} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h. \quad (13)$$

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

Table 1. The number of plots sampled from each zone in seals.data

Zone	Number of Plots	Plots Sampled
1	68	17
2	84	12
3	48	11
Total	200	40

The variances of the estimates are

$$\hat{V}(\hat{t}_{\text{str}}) = \sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) N_h^2 \frac{s_h^2}{n_h}, \quad (14)$$

$$\hat{V}(\bar{y}_{\text{str}}) = \sum_{h=1}^H \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}. \quad (15)$$

The approximate $100(1 - \alpha)\%$ confidence interval for the mean is $\bar{y}_{\text{str}} \pm z_{\alpha/2} \text{SE}(\bar{y}_{\text{str}})$.

To implement stratified sampling analysis in PROC SURVEYMEANS, create a weight of N_h/n_h for each observation in stratum h . Other than the data file with the weight value in it, a file is needed that specifies stratum name and the corresponding N_h . Specifically, N_h should take on the variable name of ‘_total_’ in this file. In PROC SURVEYMEANS, a statement specifying the name of the stratum variable should be added.

This example uses ‘seals.dat’ (Lohr, 1999, p. 123), which is on the number of breathing holes found in sampled areas of Svalbard fjords. The study was intended to estimate ringed seal populations. The study area was divided into three zones which define the strata. The total number of plots and the plots sampled in each zone are presented in Table 1.

The number of breathing holes in each sampled plot was recorded. Lohr (1999) showed the data as an example of post-stratification, where an SRS was taken from the entire population and then the number of units belonging to each stratum in the selected sample was recorded. For illustration purposes, these data are treated as a regular stratified sampling example because of its simplicity and accessibility. To estimate the total number of breathing holes in the study region, along with its standard error, we use the following SAS program:

```

data totals; /* This file gives Nh for each stratum */
input zone _total_ @@;
cards;
1 68 2 84 3 48
;
run;

data seals;
infile 'c:\sampling\seals.dat' dsd firstobs=2;
input zone holes;
if zone=1 then wt=68/17;
if zone=2 then wt=84/12;
if zone=3 then wt=48/11;
run;

proc surveymeans data=seals total=totals mean clm sum clsum;
    /* 'total=' specifies the name of the file containing Nh */
strata zone/list;
    /* the option 'list' gives more detailed information about
       each stratum */
var holes;
weight wt;
run;

```

This program leads to the following SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Observations	40
Sum of Weights	200

Stratum Information

Index	Stratum		Population		Sampling		N
	zone	Total	Rate	N Obs	Variable		
1	1	68	25.0%	17	holes	17	
2	2	84	14.3%	12	holes	12	
3	3	48	22.9%	11	holes	11	

Statistics

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

Variable	Mean	Std Error of Mean	Lower 95% CL for Mean	Upper 95% CL for Mean	Sum	Std Dev
Holes	4.985909	0.590132	3.790188	6.181631	997.181818	118.026447

Variable	Lower 95% CL for Sum	Upper 95% CL for Sum
holes	758.037521	1236.326115

Because the total was being estimated, examine the sum. The number of breathing holes in the study region is estimated to be 997.1818 with standard error being 118.0264.

One-Stage Cluster Sampling with an SRS of Clusters

In cluster sampling, the population is divided into blocks, called clusters or primary sampling units (psus). Individual elements, which are secondary sampling units (ssus), are allowed in the sample only if they belong to a cluster that is included in the sample. Consider one-stage cluster sampling, where every element within a sampled cluster is included in the sample. Note that this just becomes an SRS with the units being the clusters and the variable on the unit being the total for the cluster.

The notation for cluster sampling is quite different than that for SRS and stratified sampling. It is defined as follows:

N = number of clusters or psus in the population

n = number of clusters or psus included in the sample

M_i = number of ssus in the i^{th} cluster

$K = \sum_{i=1}^N M_i$ = total number of ssus in the population

t_i = total in the i^{th} cluster

There are two ways to estimate population totals and means: using unbiased estimation and using ratio estimation which is biased. Applying unbiased estimation,

$$\hat{t}_{\text{unb}} = \frac{N}{n} \sum_{i \in S} t_i, \quad (16)$$

$$\hat{\bar{y}}_{\text{unb}} = \frac{\hat{t}_{\text{unb}}}{K} \quad (17)$$

with variances estimated by

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}, \quad (18)$$

where

$$s_t^2 = \frac{1}{n-1} \sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N} \right)^2,$$

and

$$\hat{V}(\hat{\bar{y}}_{\text{unb}}) = \frac{1}{K^2} \hat{V}(\hat{t}_{\text{unb}}). \quad (19)$$

When K is not known, only ratio estimation can be applied, which we will discuss later.

To obtain the unbiased estimate of the total in SAS, ignore the individual elements, and use the results for the sum in simple random sampling with the cluster totals as the observations and with weight being N/n . The unbiased estimate of population mean can be obtained in a similar way through scaling the cluster totals by $1/K$.

The Green Globules data set (Lohr, 1999, p. 172) is used to demonstrate the SAS computation. The data set was originally a two-stage cluster sampling with SRS at each stage. Modifications to the data were made so that a one-stage cluster sampling scenario could be applied. Suppose the new candy Green Globules is being test marketed in an area of upstate New York. The market research firm decides to sample 6 of the 45 towns in the area and examine the number of cases of Green Globules sold in all supermarkets in the selected towns. The data set consists of two variables: town, which refers to the town the examined supermarket belongs to; and ncases, which refers to the number of cases sold in the examined supermarket. Suppose the total number of supermarkets is 252. The following SAS code reads in the data set and calculates unbiased estimates of the population total and mean.

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

```
data casesold;
  input town ncases @@;
  datalines;
1 146 1 180 1 251 1 152 1 72 1 181 1 171 1 361
1 73 1 186
2 99 2 101 2 52 2 121
3 199 3 179 3 98 3 63 3 126 3 87 3 62
4 226 4 129 4 57 4 46 4 86 4 43 4 85 4 165
5 12 5 23
6 87 6 43 6 59
;
run;

proc means data=casesold nway;
  class town;
  var ncases;
  output out=tout sum=ts;
run;

data tvalue;
  set tout;
  newy=ts/252;
  wt=45/6;
run;

proc surveymeans data=tvalue total=45 sum;
  weight wt;
  var ts newy;
run;
```

As can be seen from the SAS code, the MEANS procedure was run to save the cluster totals into an output file “tout”. The next step determines the weight, and defines the new variable “newy” by dividing the cluster totals by the total number of supermarkets, which is for the purpose of estimating the population mean. The SURVEYMEANS procedure requested the sum and produced the following output:

The SURVEYMEANS Procedure

Data Summary

Number of Observations	6
Sum of Weights	45

Statistics

Variable	Sum	Std Dev
ts	29565	10316
newy	117.321429	40.934698

The unbiased estimate of the population total (i.e., the total number of cases of Green Globules sold in the area) is 29565 with a SE of 10316. The estimate of the mean (i.e., the average number of cases of Green Globules sold in a supermarket in the area) is 117.3214 with a SE of 40.9347.

Using ratio estimation, the population mean and total are estimated by

$$\hat{\bar{y}} = \frac{\sum_{i \in S} t_i}{\sum_{i \in S} M_i}, \quad (20)$$

$$\hat{t}_r = K \hat{\bar{y}}_r, \quad (21)$$

with variances estimated by

$$\hat{V}(\hat{\bar{y}}_r) = \left(1 - \frac{n}{N}\right) \frac{1}{n \bar{M}_U^2} \frac{\sum_{i \in S} M_i^2 (\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1}, \quad (22)$$

$$\hat{V}(\hat{t}_r) = N^2 \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{\sum_{i \in S} M_i^2 (\bar{y}_i - \hat{\bar{y}}_r)^2}{n-1} \quad (23)$$

In SAS, running PROC SURVEYMEANS directly with the cluster statement on the variable of interest produces the correct analysis for ratio estimation of the mean. Note however that choosing the sum option in the surveymeans procedure does not produce an estimate of the total.

```
proc surveymeans data=casesold total=45 mean;
  cluster town;
  var ncases;
run;
```

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

The SURVEYMEANS Procedure

Data Summary

Number of Clusters	6
Number of Observations	33

Statistics

Variable	Mean	Std Error of Mean
ncases	119.454545	21.384999

Note that the ratio estimate of the mean (i.e., 119.4545) takes a different value from the unbiased estimator, and in this example it has a smaller standard error of 21.3850.

Two-Stage Cluster Sampling with SRS at Each Stage

In two-stage cluster sampling, subsample only some of the elements of selected clusters. Consider an SRS of $ssus$ is selected from each cluster. The same notations used with one-stage cluster sampling will be used, with the addition of m_i as the number of $ssus$ chosen from the i^{th} cluster.

The individual psu total is estimated by

$$\hat{t}_i = \sum_{j \in S_i} \frac{M_i}{m_i} y_{ij} = M_i \bar{y}_i \quad (24)$$

Using unbiased estimation, the population total and mean are estimated by

$$\hat{t}_{unb} = \frac{N}{n} \sum_{i \in S} \hat{t}_i, \quad (25)$$

$$\hat{\bar{y}}_{unb} = \frac{\hat{t}_{unb}}{K}. \quad (26)$$

The variances of the estimates are estimated by

$$\hat{V}(\hat{t}_{\text{unb}}) = N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n} + \frac{N}{n} \sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}, \quad (27)$$

where

$$s_t^2 = \frac{\sum_{i \in S} \left(\hat{t}_i - \frac{\hat{t}_{\text{unb}}}{N}\right)^2}{n-1}, \quad s_i^2 = \frac{\sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1},$$

and

$$\hat{V}(\hat{y}_{\text{unb}}) = \frac{1}{K^2} \hat{V}(\hat{t}_{\text{unb}}). \quad (28)$$

The unbiased estimate of the total can be obtained through the SURVEYMEANS procedure with the “cluster” option and weight of NM_i/nm_i . The variance can be seen as composed of two pieces, with the first piece being $N^2 \left(1 - \frac{n}{N}\right) \frac{s_t^2}{n}$ and the second piece being $\sum_{i \in S} \left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$:

$$\hat{V}(\hat{t}_{\text{unb}}) = \text{first piece} + \frac{N}{n} (\text{second piece})$$

The first piece of the variance is the variance of the sum obtained through running the SURVEYMEANS procedure with the “cluster” statement and a weight of NM_i/nm_i . The second piece of the variance is given by the variance of the sum through running the SAS surveymeans procedure with the cluster variable specified in the “strata” statement and a weight of M_i/m_i . The example below demonstrates the details.

The example uses the data set named ‘books.dat’ (Lohr, 1999, p. 170). A home owner with a large library needs to estimate the purchase cost and replacement value of the book collection for insurance purposes. Twelve shelves were randomly selected from a total of 44 shelves, and 5 books were randomly selected from each of the selected shelves. This is a two-stage cluster sampling with

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

SRS at each stage. In this example, $N = 44$, $n = 12$, M_i is different for each bookshelf (M_i s are given in Table 5.5 in the book), and $m_i = 5$. Suppose it is desired to estimate the total replacement value of the book collection. SAS code for manipulating the data set and conducting the analysis is as follows:

```
option ls=80 nodate;
data a;
  infile 'c:\sampling\books.dat' dlm=',' firstobs=2;
  input shelf number purchase replace;
  if shelf=2 then bigmi=26;
  if shelf=4 then bigmi=52;
  if shelf=11 then bigmi=70;
  if shelf=14 then bigmi=47;
  if shelf=20 then bigmi=5;
  if shelf=22 then bigmi=28;
  if shelf=23 then bigmi=27;
  if shelf=31 then bigmi=29;
  if shelf=37 then bigmi=21;
  if shelf=38 then bigmi=31;
  if shelf=40 then bigmi=14;
  if shelf=43 then bigmi=27;
  wt=44*bigmi/12/5;
  wt2=bigmi/5;
run;

/*to get the first piece of variance of t^ */
proc surveymeans data=a total=44 sum;
  cluster shelf;
  weight wt;
  var replace;
run;

/*to create a new file with _total_ being the cluster size Mi */
proc means data=a nway;
  class shelf;
  var bigmi;
  output out=ssize mean=_total_;
run;

/*to get the second piece of variance of t^ */
proc surveymeans data=a total=ssize sum;
  strata shelf/list;
  var replace;
  weight wt2;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Clusters	12
Number of Observations	60
Sum of Weights	1382.33333

Statistics

Variable	Sum	Std Dev
replace	32638	5613.166224

The SURVEYMEANS Procedure

Data Summary

Number of Strata	12
Number of Observations	60
Sum of Weights	377

Statistics

Variable	Sum	Std Dev
replace	8901.200000	610.297665

The first piece of the variance is 5613.166224^2 and the second piece of the variance is 610.297665^2 . Therefore,

$$\hat{t}_{\text{unb}} = 32637.73$$

$$\hat{V}(\hat{t}_{\text{unb}}) = 5613.166224^2 + \frac{44}{12} 610.297665^2 = 32873333.6$$

$$\text{SE}(\hat{t}_{\text{unb}}) = \sqrt{\hat{V}(\hat{t}_{\text{unb}})} = 5733.53$$

Again, ratio estimation could also be used with the population mean and total estimated by

$$\hat{\bar{y}}_r = \frac{\sum_{i \in S} \hat{t}_i}{\sum_{i \in S} M_i}, \quad (29)$$

$$\hat{t}_r = K \hat{\bar{y}}_r, \quad (30)$$

with

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\bar{M}^2} \left[\left(1 - \frac{n}{N}\right) \frac{s_r^2}{n} + \frac{1}{nN} \sum_{i \in S} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i} \right], \quad (31)$$

where

$$s_r^2 = \frac{\sum_{i \in S} (M_i \bar{y}_i - M_i \hat{\bar{y}}_r)^2}{n-1}, \quad s_i^2 = \frac{\sum_{j \in S_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1},$$

and

$$\hat{V}(\hat{t}_r) = K^2 \hat{V}(\hat{\bar{y}}_r). \quad (32)$$

As with the variance of the unbiased estimator, the variance of the ratio estimator can also be seen as the linear combination of two variance components. For the variance of the mean, for example, the first piece is $\frac{1}{\bar{M}^2} \left(1 - \frac{n}{N}\right) \frac{s_r^2}{n}$ and the second

piece is $\sum_{i \in S} M_i^2 \left(1 - \frac{m_i}{M_i}\right) \frac{s_i^2}{m_i}$, so that

$$\hat{V}(\hat{\bar{y}}_r) = \text{first piece} + \frac{1}{\bar{M}^2 n N} (\text{second piece}) .$$

Note that the second piece takes exactly the same form as the second piece of the estimated variance of the unbiased total, which is given by the variance of the sum through running the SAS surveymeans procedure with the cluster variable specified in the strata statement and weight of M_i/m_i . The first piece can be seen as the sample variance of $\hat{e}_i = \frac{\hat{t}_i - M_i \hat{\bar{y}}_r}{\bar{M}}$. Therefore, the first piece can be obtained by creating this variable $\hat{e}_i$ and running the surveymeans procedure for SRS on the new variable in order to get the first component of the desired variance for the ratio estimator.

Use the previous ‘books’ for unbiased estimation in two-stage cluster sampling. To estimate the average replacement cost per book use ratio estimation:

```
/*Although the data are at ssu level, this will give Mbar as m is equal
across clusters*/
proc means mean data=a nway;
  var bigmi;
run;
```

Analysis Variable : bigmi

Mean
31.416667

The above SAS code gives the estimate of average cluster size $\bar{M} = 31.4167$. Using $\bar{M}$, the ratio estimate of the mean is

$$\hat{\bar{y}}_r = \frac{\hat{t}_{\text{unb}}}{\hat{K}} = \frac{32637.73}{31.416667 * 44} = 23.6106 .$$

The following SAS code and result lead to the variance of the ratio estimate of the mean:

```
/*The new variable gettihat is created here to facilitate getting ti's
in the next step */
data a;
  set a;
  gettihat=replace*bigmi/5;
```

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

```
run;

proc means data=a nway sum mean;
  class shelf;
  var gettihat bigmi;
  output out=ratio1 sum=tihat temp1 mean=temp2 bigmi;
run;

data ratio2;
  set ratio1;
  ei=(tihat-23.61061*bigmi)/31.416667;
  keep shelf tihat bigmi ei;
run;

proc surveymeans data=ratio2 total=44 mean;
  var ei;
run;
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations   12

Statistics

Variable	Mean	Std Error of Mean
ei	7.9575595E-8	5.410291

Therefore

$$\hat{V}(\hat{\bar{y}}_r) = 5.410291^2 + \frac{1}{12 * 44 * 31.416667^2} 610.297665^2 = 29.9860$$

$$SE(\hat{\bar{y}}) = 5.4760$$

Stratified Sampling with One-Stage Cluster Sampling (Using SRS) within Each Stratum

Consider stratified sampling with one-stage cluster sampling within each stratum. Let H denote the number of strata; N_h denote the total number of psus (i.e., clusters) within stratum h , and n_h denote the selected number of psus (or clusters) within

stratum h ; t_{hi} denote the total for psu (cluster) i in stratum h , and t_h denote the total for stratum h ; M_{hi} denote the number of individual observations (ssus) for cluster i within stratum h ; s_{th}^2 denote the sample variance of t_{hi} in stratum h ; K denote the total number of individual observations.

Horvitz-Thompson estimation

Using the Horvitz-Thompson procedure (Horvitz & Thompson, 1952), the population total and mean are estimated by:

$$\hat{t}_{HT} = \sum_h \frac{N_h}{n_h} \sum_{i \in S} t_{hi} , \quad (33)$$

$$\hat{V}(\hat{t}_{HT}) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{th}^2}{n_h} . \quad (34)$$

When K is known,

$$\bar{y}_U = \frac{1}{K} \hat{t}_{HT} , \quad (35)$$

$$\hat{V}(\hat{\bar{y}}_U) = \frac{1}{K^2} \hat{V}(\hat{t}_{HT}) . \quad (36)$$

The implementation of cluster sampling under stratified sampling is straightforward in SAS. PROC SURVEYMEANS can be run with the stratum and cluster statements specified and the weight being N_h/n_h .

Table 2. Sampling design information for the ice cream spending data set

Grade	Number of Study Groups	Number of Students
7	608	1,824
8	252	1,025
9	403	1,151
Total	1,263	4,000

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

The data set used for illustration purpose comes from Example 96.1 of the SAS/STAT® 13.1 User's Guide (SAS Institute Inc., 2013). The study population is a junior high school with a total of 4,000 students in grades 7, 8, and 9. The variable of interest is how much these students spend weekly for ice cream. The clusters are study groups within grades. In each grade, a single stage cluster sampling is applied. Suppose the objective is to estimate t = the total amount spent by students (over all grades) and $\bar{y}_U$ = the mean spending per student, assuming that there are 4,000 total students. Table 2 shows the number of study groups and number of students in each grade.

Below is the SAS code that reads in the data set and conducts the analysis. Note that the variable StudyGroup identifies a student's study group. It is possible for students from different grades to have the same study group number because study groups are sequentially numbered within each grade.

```
option ls=80 nodate;
data IceCreamStudy;
  input Grade StudyGroup Spending @@;
  datalines;
7 34 7 7 34 7 7 412 4 9 27 14
7 34 2 9 230 15 9 27 15 7 501 2
9 230 8 9 230 7 7 501 3 8 59 20
7 403 4 7 403 11 8 59 13 8 59 17
8 143 12 8 143 16 8 59 18 9 235 9
8 143 10 9 312 8 9 235 6 9 235 11
9 312 10 7 321 6 8 156 19 8 156 14
7 321 3 7 321 12 7 489 2 7 489 9
7 78 1 7 78 10 7 489 2 7 156 1
7 78 6 7 412 6 7 156 2 9 301 8
;
run;

data StudyGroups;
  input Grade _total_; datalines;
7 608
8 252
9 403
;

data withweight;
  set IceCreamStudy;
  if grade=7 then wt=608/8;
  if grade=8 then wt=252/3;
  if grade=9 then wt=403/5;
run;
```

```

proc surveymeans sum data=withweight total=StudyGroups;
  strata Grade /list;
  cluster StudyGroup;
  var spending;
  weight wt;
run;

```

SAS output is presented below:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N	Clusters
1	7	608	1.32%	20	Spending	20	8
2	8	252	1.19%	9	Spending	9	3
3	9	403	1.24%	11	Spending	11	5

Statistics

Variable	Sum	Std Dev
Spending	28223	3456.556840

Therefore, $\hat{t}_{HT} = 28223$ and $SE(\hat{t}_{HT}) = \sqrt{\hat{V}(\hat{t}_{HT})} = 3456.5568$. Further, $K = 4000$, $\bar{y}_U = \frac{1}{K}\hat{t}_{HT} = 7.0557$ and $SE(\hat{y}_U) = \frac{1}{K}SE(\hat{t}_{HT}) = 0.8641$.

Ratio estimation

As before, ratio estimation can also be carried out.

The ratio estimator of the mean is determined by $\hat{y}_r = \frac{\hat{t}}{\hat{K}}$. Although $\hat{t}$ is available from the demonstration of Horvitz-Thompson estimation in the early part of this section, $\hat{K}$ needs to be obtained to be able to compute the ratio estimate. $\hat{K}$ can be obtained in a similar way as $\hat{t}$ is obtained.

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

```
data withweight;
  set IceCreamStudy;
  size=1;
  if grade=7 then wt=608/8;
  if grade=8 then wt=252/3;
  if grade=9 then wt=403/5;
run;

data StudyGroups;
  input Grade _total_ ; datalines;
  7 608
  8 252
  9 403
  ;

proc surveymeans sum data=withweight total=StudyGroups;
  strata Grade /list;
  cluster StudyGroup;
  var size;
  weight wt;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

Number of Strata	3
Number of Clusters	16
Number of Observations	40
Sum of Weights	3162.6

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N	Clusters
1	7	608	1.32%	20	size	20	8
2	8	252	1.19%	9	size	9	3
3	9	403	1.24%	11	size	11	5

Statistics

Variable	Sum	Std Dev
size	3162.600000	237.486276

From the SAS output, $\hat{K} = 3162.6000$ [and $SE(\hat{K}) = 237.4863$], which leads to the ratio estimate of $\hat{\bar{y}}_r = \frac{\hat{t}}{\hat{K}} = \frac{28223}{3162.6} = 8.9239$.

There are several approaches to obtain the variance. One approach defines the variance of $\hat{\bar{y}}_r$ as

$$\hat{V}(\hat{\bar{y}}_r) = \sum_{h=1}^H N_h^2 \frac{N_h - n_h}{N_h n_h} s_e^2, \quad (37)$$

where $e_i = \frac{t_{hi} - M_{hi} \hat{\bar{y}}_r}{\hat{K}}$.

The following SAS code creates the variable e_i and computes its sample variance, and hence the variance of $\hat{\bar{y}}_r$.

```
proc means sum n nway data=IceCreamStudy;
  class grade studygroup;
  var spending;
  output out=output1 sum=thi n=mhi;
run;

data ratio;
  set output1;
  if grade=7 then wt=608/8;
  if grade=8 then wt=252/3;
  if grade=9 then wt=403/5;
  ei=(thi-mhi*8.923860115)/3162.6;
run;

proc surveymeans sum varsum data=ratio total=StudyGroups;
  strata Grade /list;
  var ei;
  weight wt;
run;
```

SAS output:

The SURVEYMEANS Procedure

Data Summary

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

Number of Strata	3
Number of Observations	16
Sum of Weights	1263

Stratum Information

Stratum Index	Grade	Population Total	Sampling Rate	N Obs	Variable	N
1	7	608	1.32%	8	ei	8
2	8	252	1.19%	3	ei	3
3	9	403	1.24%	5	ei	5

Statistics

Variable	Sum	Std Dev	Var of Sum
ei	9.517492E-11	0.650859	0.423618

Therefore $\hat{V}(\hat{\bar{y}}_r) = 0.4236$.

The second approach to obtain the variance of the ratio estimator makes use of Taylor's theorem (Woodruff, 1971) and specifies

$$\hat{V}(\hat{\bar{y}}_r) = \frac{1}{\hat{K}^2} \left[\hat{V}(\hat{t}) + \hat{\bar{y}}_U^2 \hat{V}(\hat{K}) - 2\hat{\bar{y}}_U \text{Cov}(\hat{t}, \hat{K}) \right] \quad (38)$$

$\hat{K}$, $\hat{V}(\hat{t})$, and $\hat{V}(\hat{K})$ were previously determined, and $\text{Cov}(\hat{t}, \hat{K})$ must be estimated. This can be obtained through

$$\text{Cov}(\hat{t}, \hat{K}) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h} \right) \frac{\text{Cov}(t_{hi}, M_{hi})}{n_h}. \quad (39)$$

Again using the Ice Cream data set, the following SAS code leads to $\text{Cov}(\hat{t}, \hat{K})$:

```

/*the second approach to get the variance for ratio estimator*/
proc corr cov data=output1;
  var thi mhi;
  by grade;
run;

```

```

data covar1;
  input Grade covar nh bighn; datalines;
  7 2.42857143 8 608
  8 17.5 3 252
  9 6.45 5 403
  ;

data covar2;
  set covar1;
  cov=bighn**2*(1-nh/bighn)*covar/nh;
run;

proc means sum data=covar2 nway;
  var cov;
run;

```

SAS output:

```

          Grade = 7
        The CORR Procedure

    2 Variables:  thi    mhi

Covariance Matrix, DF = 7

              thi              mhi
thi    37.71428571    2.42857143
mhi    2.42857143    0.28571429

```

```

          Grade = 8
        The CORR Procedure

    2 Variables:  thi    mhi

Covariance Matrix, DF = 2

              thi              mhi
thi    358.3333333    17.5000000
mhi    17.5000000    1.0000000

```

Grade = 9

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

The CORR Procedure

2 Variables: thi mhi

Covariance Matrix, DF = 4

	thi	mhi
thi	85.20000000	6.45000000
mhi	6.45000000	0.70000000

The MEANS Procedure

Analysis Variable : cov

Sum
683681.12

From the previous SAS output,

$$\hat{V}(\hat{t}) = 11947785.19, \hat{K} = 3162.6, \hat{V}(\hat{K}) = 56400$$

and

$$\text{Cov}(\hat{t}, \hat{K}) = \sum_h N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\text{Cov}(t_{hi}, M_{hi})}{n_h} = 683681.12 .$$

Therefore

$$\begin{aligned} \hat{V}(\hat{y}_r) &= \frac{1}{\hat{K}^2} \left[\hat{V}(\hat{t}) + \hat{y}_u^2 \hat{V}(\hat{K}) - 2\hat{y}_u \text{Cov}(\hat{t}, \hat{K}) \right] \\ &= \frac{1}{3162.6^2} \left[11947785.19 + 8.92386^2 * 56400 - 2 * 8.92386 * 683681.12 \right] \\ &= 0.4236 \end{aligned}$$

which is similar to what is obtained from the first approach.

One-Stage Cluster Sampling with Unequal Probabilities

Sometimes sampling using unequal probabilities can prove to be more efficient and provide more accurate estimates. Two new notations are added: π_i is the inclusion probability, the probability that the i^{th} cluster is in the sample; π_{ij} is the probability that clusters i and j are both in the sample. Because this is one-stage cluster sampling, whenever a cluster is selected, all units with a cluster are selected. The Horvitz-Thompson estimator of the population total is

$$\hat{t}_{HT} = \sum_{i \in S} \frac{t_i}{\pi_i} . \quad (40)$$

The variance can be estimated by

$$\hat{V}(\hat{t}_{HT}) = \sum_{i \in S} (1 - \pi_i) \frac{t_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} . \quad (41)$$

SAS does not provide any built-in procedures to analyze sampling with unequal probabilities. The calculation of the variance of the Horvitz-Thompson estimator needs to be programmed using the IML procedure. On the other hand, the point estimate can be obtained directly by using the SURVEYMEANS procedure with the appropriate weight specified.

The analysis of sampling with probability proportional to size (pps) is illustrated for a one-stage sample using 'agpop.dat' (Lohr, 1999, p. 437). The data are from the U.S. 1992 Census of Agriculture. Only data from the state of Alabama are used for the purpose of illustration. The relevant variables are county, acres92 (i.e., the number of acres devoted to farms in 1992), and farms92 (i.e., the number of farms in 1992). The objective is to estimate the total acres in 1992 in the state of Alabama. Although data is available for all counties, first select 6 counties with probability proportional to size. Assume there is only acreage information on the 6 sampled counties, and the number of farms is assumed to be known previously for all counties. Below is the SAS code to select a pps sample and conduct the analysis based on the pps sample.

```
option ls=80 nodate;
data a;
infile 'C:\yingl\Sampling\agpop.dat' dlm=', ' firstobs=2;
input COUNTY $25. STATE $ ACRES92 ACRES87 ACRES82 FARMS92
```

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

```

FARMS87  FARMS82  LARGE92  LARGE87  LARGE82  SMALL92
SMALLF87  SMALLF82  REGION $;
if state='AL';
  /* convert missing values */
  array numval[*] _numeric_;
  do i = 1 to dim(numval);
    if numval[i]=-99 then numval[i]=.;
  end;
keep county state farms92 acres92;
run;

/* the sum of acres92 based on the original data set gives the true
population total */
proc means mean sum n;
var farms92 acres92;
run;
/* pps option in surveyselect procedure indicates sampling with
probability proportional to size, and jtprobs gives the joint
probabilities of selection */
proc surveyselect data=a out=cout method=pps sampsize=6 jtprobs;
size farms92;
run;
proc print data=cout;
run;
proc surveymeans data=cout mean sum;
var acres92;
weight SamplingWeight;
run;

proc iml;
use cout;
read all var{acres92} into t;
read all var{SelectionProb} into pi;
read all var{JtProb_1 JtProb_2 JtProb_3 JtProb_4 JtProb_5
JtProb_6}into jprob;
close cout;
n=nrow(t);
tht=t(t)*(1/pi);
var1=0;
var2=0;
do i =1 to n-1;
var1 = var1 + (t[i]**2)*(1 - pi[i])/(pi[i]**2);
do k =i+1 to n;
var2=var2+(2*t[i]*t[k]*(jprob[i,k]-
pi[i]*pi[k])/(pi[i]*pi[k]*jprob[i,k]));
end;
end;
var1 = var1 + (t[n]**2)*(1 - pi[n])/(pi[n]**2);
var=var1+var2;

```

YING LU

```
se1=sqrt(var1);
se=sqrt(var);
print tht sel se var1 var2 var;
run;
```

SAS output:

The MEANS Procedure

Variable	Mean	Sum	N
FARMS92	565.7462687	37905.00	67
ACRES92	126131.69	8450823.00	67

The SURVEYSELECT Procedure

Selection Method PPS, Without Replacement
Size Measure FARMS92

Input Data Set A
Random Number Seed 45721
Sample Size 6
Output Data Set COUT

		Selection Sampling			Prob	Weight	Unit
Obs	COUNTY	STATE	ACRES92	FARMS92			
1	LOWNDES COUNTY	AL	199714	315	0.04986	20.0556	1
2	WASHINGTON COUNTY	AL	85086	361	0.05714	17.5000	2
3	ETOWAH COUNTY	AL	85821	774	0.12252	8.1621	3
4	LIMESTONE COUNTY	AL	207226	910	0.14404	6.9423	4
5	BALDWIN COUNTY	AL	167832	941	0.14895	6.7136	5
6	LAUDERDALE COUNTY	AL	201892	1143	0.18093	5.5271	6

Obs	JtProb_1	JtProb_2	JtProb_3	JtProb_4	JtProb_5	JtProb_6
1	0.000000	0.002414	0.005175	0.006085	0.006292	0.007579
2	0.002414	0.000000	0.005940	0.006984	0.007222	0.008699
3	0.005175	0.005940	0.000000	0.015307	0.015829	0.019066
4	0.006085	0.006984	0.015307	0.000000	0.018849	0.022702
5	0.006292	0.007222	0.015829	0.018849	0.000000	0.023569
6	0.007579	0.008700	0.019066	0.022702	0.023569	0.000000

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

The SURVEYMEANS Procedure

Data Summary

Number of Observations	6
Sum of Weights	64.9007321

Statistics

Variable	Mean	Std Error of Mean	Sum	Std Dev
ACRES92	152173	28311	9876129	2914572

THT	SE1	SE	VAR1	VAR2	VAR
9876129.3	4651452.5	2934751.8	2.1636E13	-1.302E13	8.6128E12

Therefore

$$\hat{t}_{HT} = 9876129.3, \text{SE}(\hat{t}_{HT}) = 2934751.8 .$$

General Complex Surveys

General complex surveys can usually be analyzed using Horvitz-Thompson estimation or ratio estimation given that the probability and joint probability of selection are obtainable. Given the extensive analyses needed for complex survey designs, SAS implementation is not given here. In general, when a survey involves several stages of cluster and stratified sampling, methodologies illustrated above can be applied stage-by-stage to conduct analyses from the bottom stage to the top stage.

For general illustration, consider a data set (Lohr, 1999) that relates to a 1991 nationwide survey conducted in the Gambia designed to estimate the prevalence of bed net use in rural areas.

“The sampling frame consisted of all rural villages of fewer than 3000 people in The Gambia. The villages were stratified by three geographic regions (eastern, central, and western) and by whether the village had a public health clinic (PHC) or not. In each region five districts were chosen with probability proportional to the district population as estimated in the

1983 national census. In each district four villages were chosen, again with probability proportional to census population: two PHC villages and two non-PHC villages. Finally, six compounds were chosen more or less randomly from each village, and a researcher recorded the number of beds and nets, along with other information, for each compound.” (p. 223)

The two variables of interest are Y = the number of beds with nets and X = the number of beds, and the sampling scheme used is

1. stratified sampling (region)
2. cluster sampling (district)
3. stratified sampling (PHC/non-PHC)
4. cluster sampling (village)
5. SRS (compound)

Let $\hat{t}_y$ denote the estimated total for number of beds with nets, $\hat{t}_x$ denote the estimated total for number of beds, M denote the total number of clusters, and m denote the number of clusters selected. Notations for subscripts are as follows: Let r denote region ($r = 1, 2, 3$), rd denote district d in region r , rdp denote PHC situation p (with 1 indicating PHC and 2 indicating non-PHC) in district d in region r , $rdpv$ denotes village v with PHC situation p in district d and in region r , and $rdpvc$ denotes compound c in village v with PHC situation p in district d and in region r .

Horvitz-Thompson estimation of the total

Let π_i be the inclusion probability, the probability that the i^{th} unit is in the sample; and π_{ij} is the probability that units i and j are both in the sample. The Horvitz-Thompson procedure specifies that the calculation of the estimates or standard errors should start from the bottom stage up. During each stage of sampling,

$$\hat{t}_{HT} = \sum_{i \in S} \frac{\hat{t}_i}{\pi_i} \quad (42)$$

with variance of

$$\hat{V}(\hat{t}_{HT}) = \sum_{i \in S} (1 - \pi_i) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in S} \sum_{\substack{k \in S \\ k \neq i}} \frac{\pi_{ik} - \pi_i \pi_k}{\pi_{ik}} \frac{t_i}{\pi_i} \frac{t_k}{\pi_k} + \sum_{i \in S} \frac{\hat{V}(\hat{t}_i)}{\pi_i}, \quad (43)$$

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

where $\hat{t}_i$ and $\hat{V}(\hat{t}_i)$ are obtained from the sampling analysis at the lower stage.

The steps below give $\hat{t}y$ and $\hat{V}(\hat{t}y)$, and the same procedure would be followed to get $\hat{t}x$ and $\hat{V}(\hat{t}x)$. At the 5th stage, which is the lowest stage,

$$\hat{t}y_{rdpv} = \frac{M_{rdpv}}{6} \sum_{c=1}^6 y_{rdpvc} \quad , \quad (44)$$

$$\hat{V}[\hat{t}y_{rdpv}] = M_{rdpv}^2 \frac{M_{rdpv} - 6}{6M_{rdpv}} sy_{rdpvc}^2 \quad . \quad (45)$$

At the 4th stage,

$$\hat{t}y_{rdp} = \sum_{v=1}^2 \frac{\hat{t}y_{rdpv}}{\pi_v} \quad , \quad (46)$$

where $\pi_v = \frac{POP_{rdpv}}{POP_{rdp}} \cdot 2$, which is the inclusion probability for village v , and which is proportional to census population.

$$\hat{V}[\hat{t}y_{rdp}] = \sum_{v=1}^2 (1 - \pi_v) \frac{\hat{t}y_{rdpv}^2}{\pi_v^2} + \sum_{v=1}^2 \sum_{k \neq v}^2 \frac{\pi_{vk} - \pi_v \pi_k}{\pi_{vk}} \frac{\hat{t}y_{rdpv} \hat{t}y_{rdpk}}{\pi_v \pi_k} + \sum_{v=1}^2 \frac{\hat{V}[\hat{t}y_{rdpv}]}{\pi_v} \quad , \quad (47)$$

where π_{vk} is the inclusion probability for both village v and village k .

At the 3rd stage:

$$\hat{t}y_{rd} = \sum_{p=1}^2 \hat{t}y_{rdp} \quad , \quad (48)$$

$$\hat{V}[\hat{t}y_{rd}] = \sum_{p=1}^2 \hat{V}[\hat{t}y_{rdp}] \quad . \quad (49)$$

At the 2nd stage:

$$\hat{t}y_r = \sum_{d=1}^5 \frac{\hat{t}y_{rd}}{\pi_d}, \quad (50)$$

where $\pi_d = \frac{POP_{rd}}{POP_r} \cdot 5$, which is the inclusion probability for district d , and which is proportional to census population.

$$\hat{V}[\hat{t}y_r] = \sum_{d=1}^5 (1 - \pi_d) \frac{\hat{t}y_{rd}^2}{\pi_d^2} + \sum_{d=1}^5 \sum_{k \neq d} \frac{\pi_{dk} - \pi_d \pi_k}{\pi_{dk}} \frac{\hat{t}y_{rd} \hat{t}y_{rk}}{\pi_d \pi_k} + \sum_{d=1}^5 \frac{\hat{V}[\hat{t}y_{rd}]}{\pi_d}, \quad (51)$$

where π_{dk} is the inclusion probability for both village v and village k .

At the top stage:

$$\hat{t}y = \sum_{r=1}^3 \hat{t}y_r, \quad (52)$$

$$\hat{V}[\hat{t}y] = \sum_{r=1}^3 \hat{V}[\hat{t}y_r]. \quad (53)$$

Ratio estimation of the mean

To estimate the population mean, ratio estimation can usually be used at the top/first level of the sampling design. The general formula for ratio estimation is as follows:

$$\hat{B} = \frac{\hat{t}y}{\hat{t}x}, \quad (54)$$

$$\hat{V}(\hat{B}) = \frac{\hat{V}(\hat{t}y - \hat{B}\hat{t}x)}{\hat{t}x^2} = \frac{1}{\hat{t}x^2} \sum_{r=1}^3 \hat{V}(\hat{t}y_r - \hat{B}\hat{t}x_r). \quad (55)$$

To compute $\hat{V}(\hat{B})$, create a new variable $e_i = y_i - \hat{B}x_i$; then the estimation of $\hat{V}(\hat{B})$ reduces to the estimation of $\hat{V}(\hat{B}) = \frac{1}{\hat{t}x^2} \hat{V}[\hat{t}e]$. The estimation of $\hat{V}[\hat{t}e]$ follows a similar procedure as the estimation of $\hat{V}[\hat{t}y]$ as illustrated above.

ANALYZING DIFFERENT SAMPLING DESIGNS (SAS)

Specifically, using the same example, obtain $\hat{V}[\hat{te}]$ by conducting stage-by-stage analysis.

At the 5th stage, which is the lowest stage,

$$\hat{te}_{rdpv} = \frac{M_{rdpv}}{6} \sum_{c=1}^6 e_{rdpvc} = \hat{ty}_{rdpv} - \hat{B}\hat{tx}_{rdpv} , \quad (56)$$

$$\hat{V}[\hat{te}_{rdpv}] = M_{rdpv}^2 \frac{M_{rdpv} - 6}{6M_{rdpv}} se_{rdpvc}^2 . \quad (57)$$

At the 4th stage:

$$\hat{te}_{rdp} = \sum_{v=1}^2 \frac{\hat{te}_{rdpv}}{\pi_v} = \hat{ty}_{rdp} - \hat{B}\hat{tx}_{rdp} , \quad (58)$$

where $\pi_v = \frac{POP_{rdpv}}{POP_{rdp}} \cdot 2$, which is the inclusion probability for village v , and which is proportional to census population.

$$\hat{V}[\hat{te}_{rdp}] = \sum_{v=1}^2 (1 - \pi_v) \frac{\hat{ty}_{rdpv}^2}{\pi_v^2} + \sum_{v=1}^2 \sum_{k \neq v} \frac{\pi_{vk} - \pi_v \pi_k}{\pi_{vk}} \frac{\hat{te}_{rdpv} \hat{te}_{rdpk}}{\pi_v \pi_k} + \sum_{v=1}^2 \frac{\hat{V}[\hat{te}_{rdpv}]}{\pi_v} , \quad (59)$$

where π_{vk} is the inclusion probability for both village v and village k .

At the 3rd stage:

$$\hat{te}_{rd} = \sum_{p=1}^2 \hat{te}_{rdp} = \hat{ty}_{rd} - \hat{B}\hat{tx}_{rd} , \quad (60)$$

$$\hat{V}[\hat{te}_{rd}] = \sum_{p=1}^2 \hat{V}[\hat{te}_{rdp}] . \quad (61)$$

At the 2nd stage:

$$\hat{te}_r = \sum_{d=1}^5 \frac{\hat{te}_{rd}}{\pi_d} = \hat{ty}_{rd} - \hat{B}\hat{tx}_{rd}, \quad (62)$$

where $\pi_d = \frac{POP_{rd}}{POP_r} \cdot 5$, which is the inclusion probability for district d , and which is proportional to census population.

$$\hat{V}[\hat{te}_r] = \sum_{d=1}^5 (1 - \pi_d) \frac{\hat{te}_{rd}^2}{\pi_d^2} + \sum_{d=1}^5 \sum_{k \neq d} \frac{\pi_{dk} - \pi_d \pi_k}{\pi_{dk}} \frac{\hat{te}_{rd} \hat{te}_{rk}}{\pi_d \pi_k} + \sum_{d=1}^5 \frac{\hat{V}[\hat{te}_{rd}]}{\pi_d}, \quad (63)$$

where π_{dk} is the inclusion probability for both village v and village k .

At the top stage:

$$\hat{te} = \sum_{r=1}^3 \hat{te}_r, \quad (64)$$

$$\hat{V}[\hat{te}] = \sum_{r=1}^3 \hat{V}[\hat{te}_r]. \quad (65)$$

Conclusion

Although complex sampling schemes could effectively reduce the time and financial resources required to estimate population characteristics in education, such sampling might also introduce error and bias if the data resulting from the sample designs were analyzed in an inappropriate way. With very few established software systems available to conduct analyses for complex sampling designs in a “point-and-click” way, it would be helpful to have reference documentation that gives instructions on how to use SAS to conduct sampling analyses by utilizing the current features associated with the SURVEYMEANS procedure as well as the SAS IML programming environment. This paper reviewed various sampling designs within the framework of probability sampling and provided documentation on how to use SAS to estimate means and proportions in different sampling designs.

Acknowledgements

The author is grateful to Professor John P. Buonaccorsi. His lectures on survey sampling helped shape the focus of this study. He also provided insightful comments and suggestions on an earlier version of this manuscript.

References

- Horvitz, D. G., & Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260), 663-685. doi: [10.1080/01621459.1952.10483446](https://doi.org/10.1080/01621459.1952.10483446)
- Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Duxbury Press.
- SAS Institute Inc. (2013). *SAS/STAT® 13.1 user's guide*. Cary, NC: SAS Institute.
- Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66(334), 411-414. doi: [10.1080/01621459.1971.10482279](https://doi.org/10.1080/01621459.1971.10482279)

Emerging Scholars Graphing Effects as Fuzzy Numbers In Meta-Analysis

Christopher G. Thompson

Florida State University
Tallahassee, FL

Prior to quantitative analyses, meta-analysts often explore descriptive characteristics of effect sizes. A graphic is proposed that treats effect sizes as fuzzy numbers. This plot can provide meta-analysts with such information such as heterogeneity of effects, precision of estimates, possible clusters, and existence of outliers.

Keywords: Meta-analysis, fuzzy numbers, meta-analysis graphics

---

Meta-Analysis and Graphics

Meta-analysis is the statistical science of analyzing a collection of results from a set of studies with the intention of integrating individual findings (Glass, 1976). Over the past several decades, many fields have not only embraced the practical uses of meta-analysis, but have consistently strived to explore, enhance, and create new methodologies to answer complex research questions. Graphical displays of data in meta-analysis are intrinsic to answering such questions. As meta-analysis has evolved as a science, several graphical approaches have been developed (for overviews and usage suggestions, see Anzures-Cabrera & Higgins, 2010; Bax et al., 2009). The study and introduction of new graphical methods remains active today (e.g., Schild & Voracek, 2015).

The purposes of these graphics vary. Some aim to explore effect-size heterogeneity. Others reveal the possibility of publication bias. Perhaps the most widely used graphic is the forest plot. A forest plot vertically or horizontally “stacks” confidence intervals of collected effects in some predetermined order (e.g., by date, alphabetically). This display has the potential to inform meta-analysts on a variety

---

Christopher Thompson is a Ph.D Candidate in Measurement and Statistics in the Department of Educational Psychology and Learning Systems. Email him at: cgthompson@fsu.edu.

of descriptive qualities that are critical to interpretation including within-study variability, study-to-study variability, presence of outliers, and even differential effects due to moderators. However, forest plots have their weaknesses. Because forest plots stack effect-size confidence intervals, plots with a large number of effects become quite large, and thus can suffer from decreased usability and interpretability. Related to this, large forest plots may cover several pages, further complicating efficient and precise interpretation. As will be shown later, the fuzzy number graphic approach contrasts this by superimposing effects.

When vertically or horizontally stacking effects using forest plots, one must make a decision as to how to order effects. Some orderings can lead to faulty inferences (e.g., false suggestions of effect-size clusters). For example, in some meta-analyses, ordering effects by moderators (e.g., by publication year) may help to explore possible or suspected differential effects. In other meta-analyses, random or non-meaningful effect-size orderings (e.g., by author name) have the potential to erroneously suggest moderating effects or effect-size clusters. The proposed plot automatically orders effects according to magnitude.

The purpose of this article is to describe a new method of graphing effect sizes in meta-analysis using the same study information required for forest plots (i.e., estimates of effect sizes and their variances). The fuzzy number plot may prove useful when attempting to initially describe a collection of effects. To begin, condensed overviews of fuzzy sets and fuzzy-numbers are provided. These central concepts are then developed for the context of meta-analysis. Several examples using existing meta-analyses are provided. The paper concludes with several remarks.

Crisp Sets and Fuzzy Sets

In classical set theory, at the most intuitive level, a set can be described as a collection of elements (Halmos, 1960). An element is either in a set or it is not. In the context of fuzzy set theory, classical sets are often referred to as crisp sets. This terminology is derived from the indicator-like nature of the membership function which defines a crisp set. Let F be the crisp set of elements x from some universal set X . The membership function of X , denoted as μ_F , assigns a membership grade to all elements $x \in X$. In the case of crisp sets, membership functions are deterministic, namely

$$\mu_F = \begin{cases} 1, & x \in F \\ 0, & x \notin F \end{cases} . \quad (1)$$

As shown in (1), an element x is either completely included in or excluded from F .

Moving away from crisp sets, a fuzzy set $\tilde{F}$ containing elements $x \in X$ is also defined by a membership function, which can be presented as a mapping of X to the closed interval $[0, 1]$. In the case of crisp sets, the membership function is a mapping of X to the finite set $\{0, 1\}$. The distinction between mapping to an interval of membership grades and mapping to a finite set of membership grades is critical. The first mapping assigns a membership grade to each element $x \in X$ from the closed interval $[0, 1]$ while the second mapping assigns a membership grade to each element $x \in X$ from the finite set $\{0, 1\}$. Succinctly, a fuzzy set $\tilde{F}$ can be expressed as

$$\tilde{F} = \{(x, \mu_{\tilde{F}}(x)) \mid x \in X, \mu_{\tilde{F}}(x) \in [0, 1]\} . \quad (2)$$

As $\mu_{\tilde{F}}(x)$ approaches unity, the degree of membership of x in $\tilde{F}$ increases, and as $\mu_{\tilde{F}}(x)$ approaches zero, the degree of membership of x in $\tilde{F}$ decreases (Zadeh, 1965). A membership grade of unity implies an element is completely included in the fuzzy set, while a membership grade of zero implies the element is completely excluded from the fuzzy set. One can consider membership grades for fuzzy sets as numerical specifications as to how well some element $x \in X$ agrees with the imprecise mechanism (the membership function) which formulates the fuzzy set (Negoiță & Ralescu, 1987). Below are two heuristic examples of fuzzy sets.

Magnetic Strength Example

Suppose $\tilde{A}$ is a fuzzy set of strong magnetic field strengths (Gs). What precisely determines a strong magnetic field is a fuzzy concept. In this example, $\mu_{\tilde{A}}$ assigns membership grades to elements a in $\tilde{A}$ such that larger membership grades correspond with stronger magnetic strengths (i.e., greater Gs). As an example, one possible fuzzy set consisting of four elements is $\tilde{A} = \{(30, 0.2), (74, 0.6), (96, 0.7), (302, 1)\}$. For the fuzzy set $\tilde{A}$, an element (δ, ϵ) is the pairing of a magnetic strength (δ) with its membership grade (ϵ) . Although the explicit membership function is not presented here, it is clearly evident that as magnetic strength increases, so does the membership grade.

Child Intelligence Example

Suppose $\tilde{B}$ is a fuzzy set of high-scoring results from the Wechsler Preschool and Primary Scale of Intelligence (Wechsler, 2002). The concept of a high-scoring result is also fuzzy. As with the previous example, $\mu_{\tilde{B}}$ assigns membership grades to elements b in $\tilde{B}$. Larger membership grades correspond with higher intelligence scores. Furthermore, suppose that $\tilde{B} = \{(61, 0.2), (81, 0.3), (111, 0.8), (145, 1)\}$. Similar to the previous example, for the fuzzy set $\tilde{B}$, an element (δ, ϵ) is the pairing of an intelligence score (δ) with its membership grade (ϵ). As a child's intelligence score increases, the grade of membership increases.

Confidence Intervals as Fuzzy Sets

One relevant application of fuzzy sets in meta-analysis uses information from effect-size confidence intervals. The very nature of the confidence interval and its underlying notion of precision of estimating a parameter aligns with the ability of fuzzy sets (and later on, fuzzy numbers) to express imprecise beliefs regarding set membership. The pairing of an effect-size estimate and its sample variance provide insight into the precision of an estimate. This notion can also be thought of as representing the degree of fuzziness which exists between the true population parameter (here, an effect size) and the naturally imprecise estimate. Further detail on how to use confidence interval information with fuzzy numbers is presented later.

Select Fuzzy Set Attributes

Before introducing fuzzy numbers, several basic properties of fuzzy sets must be discussed. These definitions are presented in order to move towards defining fuzzy numbers. As will be discussed later, fuzzy numbers are fuzzy sets which satisfy several specifications.

First, the height of a fuzzy set $\tilde{F}$ describes the largest membership grade, or the largest $\mu_{\tilde{F}}(x)$ value, and is denoted as $\text{hgt}(\tilde{F})$. The height of a fuzzy set will necessarily be no larger than unity and will be larger than zero. Related to this, the core of a fuzzy set $\tilde{F}$ is the crisp set of all $x \in X$ having maximal membership grade, denoted as $\text{core}(\tilde{F})$. It is critical to note that, although we are discussing properties of fuzzy sets, $\text{core}(\tilde{F})$ is a crisp set.

The α -cut (alpha cut) of a fuzzy set $\tilde{F}$, denoted as $\text{cut}_\alpha(\tilde{F})$, is the crisp set of all $x \in X$ for which $\mu_{\tilde{F}}(x) \geq \alpha$, where $\alpha \in [0, 1]$. Put another way, the α -cut is a crisp set with all elements having a membership grade greater than or equal to some value in the closed interval $[0, 1]$. When working with α -cuts from fuzzy sets in the context of statistical analysis, it should be noted that the α for determining an α -cut is completely unrelated to the α commonly used to specify Type I error. To avoid confusion, Type I error is denoted here by α' .

The last definition required to introduce the concept of fuzzy numbers is the convexity of a fuzzy set. We say that a fuzzy set $\tilde{F}$ is convex if, for $u, v \in \text{cut}_\alpha(\tilde{F})$ and all $\alpha \in [0, 1]$, it holds that

$$\phi u + (1 - \phi)v \in \text{cut}_\alpha(\tilde{F}) \quad \forall \phi \in [0, 1] . \quad (3)$$

To describe (3) in another light, a fuzzy set $\tilde{F}$ is convex if all α -cuts of $\tilde{F}$, which are themselves crisp sets, are convex (Zadeh, 1965). From here we proceed to introducing fuzzy numbers.

Fuzzy Numbers

A fuzzy number is a fuzzy set satisfying several requirements. More specifically, the fuzzy number $\tilde{f}$ may be defined from a fuzzy set $\tilde{F}$ if the following properties hold:

1. $\tilde{F}$ is convex
2. $\text{hgt}(\tilde{F}) = 1$
3. $|\text{core}(\tilde{F})| = 1$
4. $\mu_{\tilde{F}}$ is at least piecewise continuous,

where $|\cdot|$ denotes the cardinality of a crisp set. For consistency, we use lower case letters to denote fuzzy numbers and uppercase letters to denote general fuzzy sets. Just as $\tilde{F}$ was defined by its membership function $\mu_{\tilde{F}}$, the fuzzy number $\tilde{f}$ can be defined by its membership function $\mu_{\tilde{f}}$. There are infinitely many possibilities for creating $\tilde{f}$ from $\tilde{F}$ by defining the fuzzy number membership function $\mu_{\tilde{f}}$ in

Figure 1. Graphic of single triangular fuzzy number

different ways. A few specific types of fuzzy numbers are common in engineering, soft computing, and other fields. One of these common fuzzy numbers is particularly pertinent for meta-analytic applications, namely the triangular fuzzy number. Although other types of fuzzy numbers are available, using the triangular fuzzy number provides sensible comparability between the membership grade and the confidence interval. Rationale for this choice is discussed later.

A triangular fuzzy number $\tilde{f} = \text{tfn}(\gamma_l, \ddot{f}, \gamma_r)$, with modal value $\ddot{f}$ and respective left- and right-hand worst-case deviations γ_l and γ_r , is defined by the membership function for all $f \in \tilde{f}$:

$$\mu_{\tilde{f}}(f) = \begin{cases} \frac{f - \gamma_l}{\ddot{f} - \gamma_l}, & \gamma_l < f \leq \ddot{f} \\ \frac{\ddot{f} - \gamma_r}{\ddot{f} - \gamma_r}, & \ddot{f} \leq f < \gamma_r \\ 0, & f \geq \gamma_r \cup f \leq \gamma_l \end{cases} \quad (4)$$

In (4), $\ddot{f}$ is a unique element with the maximal membership grade of unity. Elements outside the bounds of the worst-case deviations have a membership grade equal to zero, and thus are not in the fuzzy set. All other elements are assigned respective membership grades from the open interval (0, 1) by (4).

Figure 1 shows a graphical representation of a single triangular fuzzy number using notation described above. The vertical axis represents the membership function $\mu_{\tilde{f}}$ while the horizontal axis represents values of $x \in X$. The modal value and worst-case deviations refer to the center and edges of the triangle. There is a symmetry about $\ddot{f}$ in Figure 1, which is a specific type of triangular fuzzy number; asymmetrical triangular fuzzy numbers are also possible.

Fuzzy Numbers for Meta-Analysis

Each primary study in a meta-analysis includes at least one measure of effect, as well as an estimate of effect-size variability. This information is required for a forest plot. I propose to use this same information to create fuzzy numbers from effect sizes as an alternative way to represent the precision of effect-size estimation. For a meta-analysis with a collection of k studies, each with an effect size (T_i) and known variance (v_i), the i^{th} triangular fuzzy number is defined as

$$\tilde{\mathcal{T}}_i = \text{tfn}(\gamma_{li}, \ddot{t}_i, \gamma_{ri}) = \text{tfn}\left(T_i - Z_{\alpha'/2} \sqrt{v_i}, T_i, T_i + Z_{\alpha'/2} \sqrt{v_i}\right), \quad (5)$$

where $Z_{\alpha'/2}$ is the critical value of the standard normal distribution with a two-tailed Type I error rate of α' and $i = 1, \dots, k$. Recall that α is used to denote and α -cut and α' to denote Type I error. In (5), the modal value of the fuzzy number is the effect-size estimate itself (T_i), and both worst-case deviations come from edges of the original confidence interval. The decreasing monotonicity moving outward in both directions from the effect-size estimate (resulting in strictly decreasing membership grades) closely resembles the inherent nature of a confidence interval, and more specifically its width. An argument for the choice of a triangular fuzzy number to represent a confidence interval has also been made by Yao, Su, and Shih (2008). Here, we use information derived using a standard normal confidence interval. The above definitions are valid for other types of confidence intervals as well.

Defining the membership function for (5), we simply revise (4) such that for all $t \in \tilde{\mathcal{T}}$

$$\begin{aligned}
\mu_{\tilde{T}_i}(t) &= \begin{cases} \frac{t-\gamma_1}{\ddot{t}-\gamma_1}, & \gamma_1 < t \leq \ddot{t} \\ \frac{t-\gamma_r}{\ddot{t}-\gamma_r}, & \ddot{t} \leq t < \gamma_r \\ 0, & t \geq \gamma_r \cup t \leq \gamma_1 \end{cases} \\
&= \begin{cases} \frac{(t)-(T_i-Z_{\alpha'/2}\sqrt{v_i})}{(T_i)-(T_i-Z_{\alpha'/2}\sqrt{v_i})}, & (T_i-Z_{\alpha'/2}\sqrt{v_i}) < t \leq T_i \\ \frac{(t)-(T_i+Z_{\alpha'/2}\sqrt{v_i})}{(T_i)-(T_i+Z_{\alpha'/2}\sqrt{v_i})}, & T_i \leq t < T_i+Z_{\alpha'/2}\sqrt{v_i} \\ 0, & t \geq T_i+Z_{\alpha'/2}\sqrt{v_i} \cup t \leq T_i-Z_{\alpha'/2}\sqrt{v_i} \end{cases} \\
&= \begin{cases} 1+\frac{t-T_i}{Z_{\alpha'/2}\sqrt{v_i}}, & \gamma_1 < t \leq \ddot{t} \\ 1-\frac{t-T_i}{Z_{\alpha'/2}\sqrt{v_i}}, & \ddot{t} \leq t < \gamma_r \\ 0, & t \geq T_i+Z_{\alpha'/2}\sqrt{v_i} \cup t \leq T_i-Z_{\alpha'/2}\sqrt{v_i} \end{cases} \quad (6)
\end{aligned}$$

Any collection of studies providing ample information to create a forest plot can necessarily be used to create a fuzzy number plot. Treating effect sizes as fuzzy numbers allows for effect-size imprecision to be viewed as fuzziness in effect-size estimation. For some effect-size metrics (e.g., the standardized mean difference), the triangular fuzzy number will be symmetric about $\ddot{t} = T_i$ because $T_i + Z_{\alpha'/2}\sqrt{v_i}$ is without bound on the set of real numbers. For other metrics (e.g., correlation coefficient), there is a possibility for asymmetry due to the natural bounds of the effect-size metric. Furthermore, it is possible to use other measures of variability to define worst-case deviations for triangular fuzzy numbers. One example would be the median absolute deviation. The use of variance in this paper is solely to coincide with confidence interval information found in forest plots.

There are obvious similarities among typical confidence intervals, probability values, and fuzzy numbers. However, the two representations of effects differ in several important ways. First, there is an intrinsic uncertainty due to randomness and uncertainty due to fuzziness. Among other things, this concerns definitions of subset domains. Fuzzy set theory replaces typical σ -algebra domains by the

universe of discourse; see Aliev, Fazlollahi, and Aliev (2004) for a more elaborate discussion. Second, a membership grade is not the same as a probability value. One reason for this distinction is that probabilities must exist in the closed interval $[0, 1]$. This is not always the case for fuzzy membership functions. Also, it need not be the case that the summation of all membership grades is unity (for a detailed discussion, see Singpurwalla & Booker, 2004). Furthermore, in the absence of fuzzy set theory (i.e., simply plotting confidence intervals in the same manner of fuzzy numbers), the metric of the vertical axis is unclear. Using fuzzy set theory, an established and interpretable membership grade is assigned to the vertical axis.

Aggregate Fuzzy Number

It may prove desirable to compute and plot some aggregate fuzzy number measure, similarly to the common practice of plotting weighted means on forest plots. One could use typical fixed-effect or random-effects weighted means as fuzzy numbers using the same formulas discussed above. Parallel to how fuzzy numbers are constructed from fuzzy effects in (5), one could use components from fixed- or random-effects confidence intervals for means. The weighted mean estimate would be the modal value and worst-case deviations would be the edges of the confidence interval for the weighted mean.

Alternatively, one could plot a mean fuzzy number $\tilde{\mathcal{T}}$ following Buckley (1985):

$$\begin{aligned}\tilde{\mathcal{T}} &= \text{tfn}(\gamma_1^m, \dot{t}^m, \gamma_r^m) \\ &= k^{-1} \bullet [\tilde{\mathcal{T}}_1 \oplus \tilde{\mathcal{T}}_2 \oplus \dots \oplus \tilde{\mathcal{T}}_k] \\ &= \text{tfn}\left(k^{-1} \sum_{i=1}^k \gamma_{li}, k^{-1} \sum_{i=1}^k \ddot{t}_i, k^{-1} \sum_{i=1}^k \gamma_{ri}\right)\end{aligned}\tag{7}$$

Where k is the number of studies in the meta-analysis, $\bullet$ is scalar multiplication and $\oplus$ is fuzzy addition (Hanss, 2005). In the context of meta-analysis, (7) can be calculated as

$$\tilde{\mathcal{T}} = \text{tfn}\left(k^{-1} \sum_{i=1}^k [T_i - Z_{\alpha'/2} \sqrt{v_i}], k^{-1} \sum_{i=1}^k T_i, k^{-1} \sum_{i=1}^k [T_i + Z_{\alpha'/2} \sqrt{v_i}]\right).\tag{8}$$

This aggregate measure is calculated as a direct function of fuzzy numbers, which is different than calculating a fuzzy number representing a fixed- or random-effects weighted mean. Typical fixed- and random-effects means in meta-analysis are inverse-variance weighted, so that studies with higher precision are afforded more weight when determining an average. The method presented in (8) averages each component of the collection of fuzzy numbers. This result is more similar to a “typical study result” than to a weighted mean. The endpoints of the fuzzy mean are not to be directly compared to those of a confidence interval. While statistical significance can be assessed using confidence intervals, this is not valid for fuzzy numbers. The fuzzy mean is essentially a representation of the fuzziness or uncertainty of a typical study.

Examples of Fuzzy Number Plots

Three examples of fuzzy number plots from published meta-analyses are shown and discussed. Forest plots with effects ordered by their magnitudes are also provided for comparison. When plotting fuzzy numbers, the degree of color shading provides a simple interpretation such that more dense (i.e., darker) shading corresponds with more fuzzy number overlap. All fuzzy number plots in this paper were produced using basic R (R Core Team, 2013) procedures along with the *FuzzyNumbers* package (Gagolewski, 2013). Forest plots were produced using the *metafor* package (Viechtbauer, 2010). R code for producing fuzzy number plots is provided in the appendix.

Exercise Training and Depressive Symptoms: A Large Number of Effects

The first example comes from a meta-analysis of the effects of exercise training on select depressive symptoms for patients with a chronic illness (Herring, Puetz, O'Connor, & Dishman, 2012). To quantify treatment effects, standardized mean differences (d) were calculated. These effects represented the mean difference between an exercise condition and a comparison condition on several mental and physical health outcomes. In total, 167 effect sizes were obtained from 90 studies. Figure 2 provides the fuzzy number plot for this data, while Figure 3 provides the forest plot for the same data.

Recall that the notion of fuzziness is represented by the vertical axis (membership grade) and the width of the triangle associated with each effect size. A wider triangle denotes a fuzzier estimate. What is immediately noticeable from Figure 2

is that the vast majority of effect-size point estimates are positive, indicating a reduction in depressive symptoms in the exercise condition (compared to the non-exercise condition). While most point estimates (i.e., circles falling on the vertical line where the membership grade is equal to unity) were positive, a dense cluster of effects falls in the interval $[0, 0.75]$. In this example, the center of the fuzzy mean $\tilde{\mathcal{J}}_1 = \text{tfn}(-0.19, -0.35, 0.89)$ is located in the approximate center of the clustered effects. Last, several large effects ($d > 1$) may be potential outliers.

This example shows how the fuzzy plot reveals some simple yet valuable descriptive features when initially describing a collection of effects. Several large effects appear to be divergent from the rest of the data, which is an indicator of possible outliers. These effects do not stand out as prominently as in Figure 3. This attribute, as well as the clustering of effects around 0 to 0.75, is more easily seen with the fuzzy number plot (Figure 2) compared to the respective forest plot (Figure 3). Furthermore, the sheer size of the forest plot is impractical. To include this forest plot in a relatively small area of a single publication page, effect sizes would need

Figure 2. Fuzzy number plot of exercise training and depressive symptoms data

Figure 3. Forest plot of exercise training and depressive symptoms data

to be graphed very close together. Consequently, the quality of interpretation can be diminished. On the other hand, increasing the size of the forest plot so that effects are not forced so close together would require more journal pages. This is not the case for the fuzzy number plot. The superimposition of effects resolves this issue.

Positive Psychology Interventions for Well-Being: The Presence of Moderators

The second example stems from a meta-analysis (Sin & Lyubomirsky, 2009) which synthesized bivariate correlations (r) from studies analyzing the effect of positive psychology interventions for well-being in depressive symptoms. This meta-analysis collected 42 effect sizes from 37 studies. The original meta-analysis included more effects and studies. Although most studies in the meta-analysis were moderately recent, some date back to the 1970s and 1980s. The example in this

Figure 4. Fuzzy number plot of psychology interventions data

paper excluded all effects before 1990 ($N = 7$). Effects were plotted using both the novel fuzzy number plot (Figure 4) and the traditional forest plot (Figure 5). This example also illustrates the capability of fuzzy number plots to explore possible moderators.

Both the fuzzy number and forest plots clearly show a preponderance of positive effects, indicating that positive psychological interventions were associated with higher states of well-being. The fuzzy number plot shows several clusters of effects. Although one of these clusters is located close to the fuzzy mean $\tilde{\mathcal{J}}_2 = \text{tfn}(-0, 0.24, 0.48)$ and shows a moderately large effect, there is also a discernible cluster of effects around zero.

To illustrate how potential moderators can be graphed on a fuzzy number plot, we again utilized the same data from Sin and Lyubomirsky (2009). One of the coded moderators in the original meta-analysis was whether physiological interventions were: 1) Individually Administered; 2) Group Administered; or 3) Self-Administered. To demonstrate the salience of this moderator, Figure 6 displays a revised fuzzy number plot with color-coded fuzzy numbers. Figure 6 suggests that those interventions administered by psychological professionals had more of an effect than interventions which were self-administered by the client.

FUZZY NUMBER PLOTS IN META-ANALYSIS

Figure 5. Forest plot of psychology interventions data

Figure 6. Fuzzy number plot of psychology interventions data with moderator

Alcohol Consumption and Coronary Heart Disease: The Presence of Outliers

The third example was taken from a published meta-analysis on alcohol consumption and its effect on select biological markers known to be associated with adult risk of coronary heart disease (Brien, Ronksley, Turner, Mukamal, & Ghali, 2011). These authors state that they “systematically reviewed the effect of experimentally manipulated alcohol consumption (alcohol use versus a period of no alcohol use) on the circulating concentrations of selected cellular and molecular biological markers of atherothrombotic conditions associated with increased coronary heart disease risk in adults without pre-existing cardiovascular disease” (Brien et al., 2011, p. 1). Analyses were reasonably extensive; over a dozen different biomarker means were examined. Here, we focus on results of one biomarker, high density lipoprotein cholesterol (HDLC).

Figure 7. Fuzzy number plot of cholesterol biomarker data

FUZZY NUMBER PLOTS IN META-ANALYSIS

Thirty-two mean differences comparing average concentration of HDLC biomarkers after alcohol consumption to the average concentration of the HDLC biomarker with no alcohol consumption were extracted. The original meta-analysis had 33 effects. However, an effect from one study had such a large variance (1.78) compared to the remaining effects, we decided that its exclusion would produce a more interpretable graphic. As with the previous example, effects were plotted using both the fuzzy number plot (Figure 7) and a traditional forest plot (Figure 8).

What is immediately noticeable from Figure 7 is the overwhelming amount of positive effect-size point estimates corresponding with a positive relationship between alcohol consumption and the HDLC biomarker. In addition, a possible outlier appears at $d = 0.63$. This extreme value also appears to have more fuzziness compared to other effects, as indicated by the larger width of the triangle. A dense cluster of effects is seen between approximately $d = 0$ and $d = 0.25$. In the approximate center of this cluster is the fuzzy mean $\tilde{\mathcal{Z}}_3 = \text{tfn}(-0.06, 0.13, 0.32)$. This graphical representation suggests a fairly weak overall effect, if any at all.

Figure 8. Forest plot of cholesterol biomarker data

Concluding Remarks

A novel method for graphing effects was proposed for describing study information in meta-analysis by way of treating and plotting effect sizes as fuzzy numbers. A brief introduction to fuzzy set theory was provided and extensions to meta-analysis were explained. Plots using data from both fuzzy number plots and forest plots were illustrated and discussed in the context of three previously published meta-analyses.

Treating effects as fuzzy numbers allows the meta-analyst to use the same information required for the common forest plot but provides several advantages. For example, plotting effects as fuzzy numbers is likely to increase usability and interpretability in situations where a large number of effects are to be meta-analyzed. In such situations, fuzzy number plots will use less publication page space than forest plots. Last, as was demonstrated, fuzzy number plots can be used to explore possible moderators and outliers at initial stages of a meta-analysis.

There are some possible limitations to the method of plotting effects as fuzzy numbers. In select instances, it may be challenging to differentiate within a group of very small effects or within a group of very large effects. Also, compared to the state-of-the-art forest plots, creating fuzzy number plots for meta-analysis involves slightly more computation on the user's end. However, R code has been provided in the appendix for this very reason.

References

- Aliev, R. A., Fazlollahi, B., & Aliev, R. R. (2004). *Soft computing and its applications in business and economics*. Studies in fuzziness and soft computing (Vol. 157). Berlin, Germany: Springer.
- Anzures-Cabrera, J., & Higgins, J. P. T. (2010). Graphical displays for meta-analysis: An overview with suggestions for practice. *Research Synthesis Methods*, 1(1), 66-80. doi: [10.1002/jrsm.6](https://doi.org/10.1002/jrsm.6)
- Bax, L., Ikeda, N., Fukui, N., Yaju, Y., Tsuruta, H., & Moons, K. G. M. (2009). More than numbers: The power of graphs in meta-analysis. *American Journal of Epidemiology*, 169(2), 249-255. doi: [10.1093/aje/kwn340](https://doi.org/10.1093/aje/kwn340)
- Brien, S., Ronksley, P., Turner, B., Mukamal, K., & Ghali, W. (2011). Effect of alcohol consumption on biological markers associated with risk of coronary heart disease: Systematic review and meta-analysis of interventional studies. *BMJ*, 342(d636). doi: [10.1136/bmj.d636](https://doi.org/10.1136/bmj.d636)

- Buckley, J. J. (1985). Ranking alternatives using fuzzy numbers. *Fuzzy Sets and Systems*, 15(1), 21-32. doi: [10.1016/0165-0114\(85\)90013-2](https://doi.org/10.1016/0165-0114(85)90013-2)
- Gagolewski, M. (2013). *Fuzzynumbers package: Tools to deal with fuzzy numbers in R* [Computer software manual]. Retrieved from <http://www.rexamine.com/resources/fuzzynumbers/>
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3-8. Available from <http://www.jstor.org/stable/1174772>
- Halmos, P. R. (1960). *Naive set theory*. New York, NY: Springer.
- Hanss, M. (2005). *Applied fuzzy arithmetic: An introduction with engineering applications*. New York, NY: Springer-Verlag.
- Herring, M. P., Puetz, T. W., O'Connor, P. J., & Dishman, R. K. (2012). Effect of exercise training on depressive symptoms among patients with a chronic illness: A systematic review and meta-analysis of randomized controlled trials. *Archives of Internal Medicine*, 172(2), 101-111. doi: [10.1001/archinternmed.2011.696](https://doi.org/10.1001/archinternmed.2011.696).
- Negoitǎ, C. V., & Ralescu, A. L. (1987). *Simulation, knowledge-based computing, and fuzzy statistics*. New York, NY: Van Nostrand Reinhold.
- R Core Team. (2013). *R: A language and environment for statistical computing* [Computer software manual]. Retrieved from <http://www.R-project.org>
- Schild, A. H. E., & Voracek, M. (2015). Finding your way out of the forest without a trail of bread crumbs: Development and evaluation of two novel displays of forest plots. *Research Synthesis Methods*, 6(1), 74-86. doi: [10.1002/jrsm.1125](https://doi.org/10.1002/jrsm.1125)
- Sin, N. L., & Lyubomirsky, S. (2009). Enhancing well-being and alleviating depressive symptoms with positive psychology interventions: A practice-friendly meta-analysis. *Journal of Clinical Psychology*, 65(5), 467-487. doi: [10.1002/jclp.20593](https://doi.org/10.1002/jclp.20593)
- Singpurwalla, N. D., & Booker, J. M. (2004). Membership functions and probability measures of fuzzy sets. *Journal of the American Statistical Association*, 99(467), 867-877. doi: [10.1198/016214504000001196](https://doi.org/10.1198/016214504000001196)
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1-48. Retrieved from <http://www.jstatsoft.org/article/view/v036i03>

Wechsler, D. (2002). *Wechsler preschool and primary scale of intelligence* (3rd ed.). San Antonio, TX: Psychological Corporation.

Yao, L. J.-S., Su, J.-S., & Shih, T.-S. (2008). Fuzzy system reliability analysis using triangular fuzzy numbers based on statistical data. *Journal of Information Science and Engineering*, 24(5), 1521-1535. Retrieved from http://www.iis.sinica.edu.tw/page/jise/2008/200809_15.html

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, 8(3), 338-353. doi: 10.1016/S0019-9958(65)90241-X

Appendix

R Code for Fuzzy Number Plot

```
library(FuzzyNumbers)

## Data is a data set with k rows (one for each effect size) and three
## columns: effect-size measures, left-hand worst case deviations, right-
## hand worst case deviations
## T    is a column of effect-size measure
## LH   is a column of left-hand worst case deviation
## RH   is a column of right-hand worst case deviation

# Color code from mages' blog:
# http://www.magesblog.com/2013/04/how-to-change-alpha-value-of-colours-
# in.html
add.alpha <- function(col, alpha=1){
  if(missing(col))
    stop("Please provide a vector of colours.")
  apply(sapply(col, col2rgb)/255, 2,
    function(x)
      rgb(x[1], x[2], x[3], alpha=alpha))}

DataM <- sum(Data$T)/length(Data$T)
DataL <- sum(Data$LH)/length(Data$T)
DataR <- sum(Data$RH)/length(Data$T)

D<-list()
for(i in 1:length(Data$T)){
  fuzzynames<-paste('a',i,sep='')
  D[[i]]<- PiecewiseLinearFuzzyNumber(Data$LH[i], Data$T[i], Data$T[i],
  Data$RH[i])}

MeanData <- PiecewiseLinearFuzzyNumber(DataL, DataM, DataM, DataR)
Dpointest<-rep(1.011,length(Data$T))
DpointestL<-rep(0,length(Data$T))
DpointestR<-rep(0,length(Data$T))

plot(D[[1]], type='l', lty=1, col=rgb(54/255, 100/255, 139/255,
alpha=.4),
      xlab='Effect Size (T)', ylab="Membership Grade")

for(i in 2:length(D)){
  plot(D[[i]], type='l', col=rgb(54/255, 100/255, 139/255, alpha=.4),
lty=1, add=TRUE)}

plot(MeanData, type='l', col='black', lwd=1, add=TRUE)
```

CHRISTOPHER THOMPSON

```
abline(h=0, lwd=1)
abline(v=0, lwd=1, lty=2, col="darkgreen")
points(x=Data$T, y=Dpointest, type="p", pch=20,
       col=add.alpha("steelblue3", alpha=0.2)[Ppointest], cex=2)
points(x=DataM, y=Dpointest[1], type="p", pch=20,
       col=add.alpha("black", alpha=0.35), cex=2)
legend(-.52, 1.1, lty=c(1,1), c("Fuzzy Numbers","Fuzzy Mean"),
      lwd=c(2,2), col=c("steelblue3", "black"), cex=.6, bty="n")
```

Instructions for Authors

Authors wishing to submit to *JMASM* may do so using the submission form at the journal's website, <http://digitalcommons.wayne.edu/jmasm>. Three areas are appropriate for *JMASM*:

1. Development or study of new statistical tests or procedures, or the comparison of existing statistical tests or procedures, using computer-intensive Monte Carlo, bootstrap, jackknife, or resampling methods;
2. Development or study of nonparametric, robust, permutation, exact, and approximate randomization methods; and
3. Applications of computer programming, preferably in Fortran (all other programming environments are welcome), related to statistical algorithms, pseudo-random number generators, simulation techniques, and self-contained executable code to carry out new or interesting statistical methods.

Elegant derivations, as well as articles with no take-home message to practitioners, have low priority. Articles based on Monte Carlo (and other computer-intensive) methods designed to evaluate new or existing techniques or practices, particularly as they relate to novel applications of modern methods to everyday data analysis problems, have high priority.

Work appearing in *Regular Articles*, *Brief Reports*, and *Emerging Scholars* is externally peer reviewed, with input from the Editorial Board; work appearing in *Statistical Software Applications and Review* and *JMASM Algorithms and Code* is internally reviewed by the Editorial Board. *JMASM* charges neither article processing fees nor submission fees.

Please observe the following guidelines when preparing manuscripts:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Articles should be submitted without a title page or abstract. There should be no material identifying authorship except in the fields of the submission form. Include a statement in the cover letter indicating that proper human subjects protocols were followed where applicable, including informed consent.
3. Manuscripts should be prepared in Microsoft Word (.doc or .docx) only (Wordperfect and .rtf formats may be acceptable – please inquire). Please note that Tex (in its various versions), Exp, and Adobe .pdf formats are designed to produce the final presentation of text. They are not amenable to the editing process, and are NOT acceptable for manuscript submission.
4. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
5. Create tables without boxes or vertical lines. Place tables, figures, and graphs "in-line", not at the end of the manuscript. Figures may be in .jpg, .tif, .png, and other formats readable by Adobe Illustrator or Photoshop.
6. The submission form requires an Abstract with a 50 word maximum, and a list of key words or phrases. Major headings are Introduction, Methodology, Results, Conclusion, and References. Center headings. Subheadings are left justified; capitalize only the first letter of each word. Sub-subheadings are left justified, indent optional.

7. Do not use underlining in the manuscript. Do not use bold, except for (a) matrices, or (b) emphasis within a table, figure, or graph. Do not number sections. Number all formulas, tables, figures, and graphs, but do not use italics, bold, or underline. Do not number references. Do not use footnotes or endnotes.
  8. In the References section, do not put quotation marks around titles of articles or books. Capitalize only the first letter of books. Italicize journal or book titles, and volume numbers. Use "&" instead of "and" in multiple author listings.
  9. Suggestions for style: Instead of "I drew a sample of 40" write "A sample of 40 was selected". Use "although" instead of "while," unless the meaning is "at the same time." Use "because" instead of "since," unless the meaning is "after." Instead of "Smith (1990) notes" write "Smith (1990) noted." Do not strike the spacebar twice after a period.
-

Journal of Modern Applied Statistical Methods

ISSN: 1538–9472

<http://digitalcommons.wayne.edu/jmasm>

PUBLISHED biannually (May, November) in partnership by:

JMASM, Inc.
PO Box 48023
Oak Park, MI 48237
ea@jmasm.com

Wayne State University Library System
Purdy Library
Detroit, MI 48202
digitalcommons@wayne.edu

Copyrights, Attribution and Usage Policies

Copyright ©2015 JMASM, Inc. *JMASM* retains the copyright for this work for the entire usual period, but grants assignors the right, after one year from the date of publication, to republish the work in whole or in part anywhere and in any format, provided reference is given to the original publication in *JMASM* (see website for further details). Readers may freely access journal content at <http://digitalcommons.wayne.edu/jmasm>.

To Advertisers

Advertisements are accepted at the discretion of the editor. Send requests for advertising information to ea@jmasm.com.