5-1-2008

Vol. 7, No. 1 (Full Issue)

JMASM Editors

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

Recommended Citation

Editors, JMASM (2008) "Vol. 7, No. 1 (Full Issue)," *Journal of Modern Applied Statistical Methods*: Vol. 7 : Iss. 1 , Article 33.

DOI: 10.22237/jmasm/1209616320

Available at: <http://digitalcommons.wayne.edu/jmasm/vol7/iss1/33>

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.

Real Data Analysis

Edited by **Shlomo S. Sawilowsky**, *Wayne State University*

A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)

The invited authors of this edited volume have been prolific in the arena of Real Data Analysis (RDA) as it applies to the social and behavioral sciences, especially in the disciplines of education and psychology. Combined, this brain trust represents 3,247 articles in refereed journals, 127 books published, US \$45.3 Million in extramural research funding, 34 teaching and 92 research awards, serve(d) as Editor/Assistant Editor/Editorial Board Member for 95 peer reviewed journals, and provide(d) ad hoc reviews for 362 journals. Their enormous footprint on real data analysis is showcased for professors, researchers, educators, administrators, and graduate students in the second text in the AERA/SIG ES Quantitative Methods series.

CONTENTS: Preface. *Shlomo S. Sawilowsky*. **PART I: FOUNDATIONS.** The Co-Evolution of Statistics and Hz, *Joseph M. Hilbe*. Effective Sample Size: A Crucial Concept, *Thomas R. Knapp*. Advances in Missing Data Methods and Implications for Educational Research, *Chao-Ying Joanne Peng, Michael Harwell, Show-Mann Liou, Lee H. Ehman*. Methods for Simulating Real World Data for the Psycho-Educational Sciences, *Todd Christopher Headrick*. How and Why I Use Real, Messy Data to Investigate Theory and Inform Decision Making, *Ted Micceri*. **PART II: STATISTICAL METHODS.** Using E-Mail Messages to Help Students Prepare for a Statistics Exam, *Schuyler Huck*. Randomization Tests: Statistical Tools for Assessing the Effects of Educational Interventions When Resources are Scarce, *Joel R. Levin*. A Skipped Multivariate Measure of Location: One- And Two-Sample Hypothesis Testing, *Rand R. Wilcox, H. J. Keselman*. Robust Step-Down Tests for Multivariate Group Differences, *Lisa M. Lix, Ian Clara, Aynslie Hinds, Charles Bernstein*. Dunn-Sidak Critical Values and *p* Values, *Roger E. Kirk, Joel Hetzer*. Controlling Experiment-wise Type I Errors: Good Advice for Simultaneous and Sequential Hypothesis Testing, *Shlomo S. Sawilowsky, Patric R. Spence*. Robustness and Power of Ordinal *d* for Paired Data, *Du Feng*. Factorial ANOVA in SPSS: Fixed-, Random-, and Mixed-Effects Models, *Richard G. Lomax, Stacy Hughey Surman*. ANOVA: Effect Sizes, Simulating Interaction vs. Main Effects, and a Modified ANOVA Table, *Shlomo S. Sawilowsky*. ANCOVA and Quasi-Experimental Design: The Legacy of Campbell and Stanley, *Shlomo S. Sawilowsky*. **PART III: MEASUREMENT:** Thinking About Item Response Theory from a Logistic Regression Perspective: A Focus on Polytomous Models, *Amery D. Wu, Bruno D. Zumbo*. Some Practical Uses of Item Response Time to Improve the Quality of Low-Stakes Achievement Test Data, *Steven L. Wise, Xiaojing Kong*. Using Moving Averages to Detect Exposed Test Items in Computer-Based Testing, *Ning Han, Ronald K. Hambleton*. An Empirical Calibration of the Effects of Multiple Sources of Measurement Error on Reliability Estimates for Individual Differences Measures, *Frank L. Schmidt, Huy Ahn Le*. Latent Structure of Attitudes toward Abortion, *C. Mitchell Dayton*. **PART IV: DATA ANALYSIS.** Hierarchical Linear Models and the Estimation of Students' Mathematics Achievement, *Kathrin A. Parks, Dudley L. Poston, Jr.* Grade Inflation: An Examination at the Institutional Level, *Sharon L. Weinberg*. Using Discrete-Time Survival Analysis to Study Gender Differences in Leaving Mathematics, *Suzanne E. Graham, Judith D. Singer*. Nonparametric procedures for testing for dropout rates on University courses with application to an Italian case study, *Rosa Arboretti Giancristofaro, Fortunato Pesarin, Luigi Salmaso, Aldo Solari*. Nonparametric Approaches for Multivariate Testing with Mixed Variables and for Ranking on Ordered Categorical Variables with an Application to the Evaluation of Ph. D. Programs, *Rosa Arboretti Giancristofaro, Fortunato Pesarin, Luigi Salmaso*. Randomized Replicated Single-case Experiments: Treatment of Pain-related Fear by Graded Exposure *In Vivo*, *Patrick Onghena, Johan W. S. Vlaeyen, Jeroen de Jong*. Whole Brain Correlations: Examining Similarity Across Conditions of Overall Patterns of Neural Activation in fMRI, *Arthur Aron, Susan Whitfield, Wemara Lichty*. Principal Component Analysis of Senate Voting Patterns. *Jan de Leeuw*

Publication Date:

2007

ISBN's:

Paperback: 978-1-59311-564-7

Hardcover: 978-1-59311-565-4

Price:

Paperback: \$39.99

Hardcover: \$73.99

Subject:

Education, Statistics

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 \$39.99 Hardcover ISBN: 978-1-59311-685-9 \$73.99

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.99

Series URL: <http://www.infoagepub.com/products/series/serlin.html>

Journal Of Modern Applied Statistical Methods

Shlomo S. Sawilowsky

Editor

College of Education
Wayne State University

Harvey Keselman

Associate Editor

Department of Psychology
University of Manitoba

Bruno D. Zumbo

Associate Editor

Measurement, Evaluation, & Research Methodology
University of British Columbia

Vance W. Berger

Assistant Editor

Biometry Research Group
National Cancer Institute

John L. Cuzzocrea

Assistant Editor

Educational Research
University of Akron

Todd C. Headrick

Assistant Editor

Educational Psychology and Special Education
Southern Illinois University-Carbondale

Alan Klockars

Assistant Editor

Educational Psychology
University of Washington

Editorial Board

Subhash Chandra Bagui
Department of Mathematics & Statistics
University of West Florida

J. Jackson Barnette
School of Public Health
University of Alabama at Birmingham

Vincent A. R. Camara
Department of Mathematics
University of South Florida

Ling Chen
Department of Statistics
Florida International University

Christopher W. Chiu
Test Development & Psychometric Rsch
Law School Admission Council, PA

Jai Won Choi
National Center for Health Statistics
Hyattsville, MD

Rahul Dhanda
Forest Pharmaceuticals
New York, NY

John N. Dyer
Dept. of Information System & Logistics
Georgia Southern University

Matthew E. Elam
Dept. of Industrial Engineering
University of Alabama

Mohammed A. El-Saidi
Accounting, Finance, Economics &
Statistics, Ferris State University

Felix Famoye
Department of Mathematics
Central Michigan University

Barbara Foster
Academic Computing Services, UT
Southwestern Medical Center, Dallas

Shiva Gautam
Department of Preventive Medicine
Vanderbilt University

Dominique Haughton
Mathematical Sciences Department
Bentley College

Scott L. Hersherberger
Department of Psychology
California State University, Long Beach

Joseph Hilbe
Departments of Statistics/ Sociology
Arizona State University

Sin-Ho Jung
Dept. of Biostatistics & Bioinformatics
Duke University

Jong-Min Kim
Statistics, Division of Science & Math
University of Minnesota

Harry Khamis
Statistical Consulting Center
Wright State University

Kallappa M. Koti
Food and Drug Administration
Rockville, MD

Tomasz J. Kozubowski
Department of Mathematics
University of Nevada

Kwan R. Lee
GlaxoSmithKline Pharmaceuticals
Collegeville, PA

Hee-Jeong Lim
Dept. of Math & Computer Science
Northern Kentucky University

Balgobin Nandram
Department of Mathematical Sciences
Worcester Polytechnic Institute

J. Sunil Rao
Dept. of Epidemiology & Biostatistics
Case Western Reserve University

Karan P. Singh
University of North Texas Health
Science Center, Fort Worth

Jianguo (Tony) Sun
Department of Statistics
University of Missouri, Columbia

Joshua M. Tebbs
Department of Statistics
Kansas State University

Dimitrios D. Thomakos
Department of Economics
Florida International University

Justin Tobias
Department of Economics
University of California-Irvine

Dawn M. VanLeeuwen
Agricultural & Extension Education
New Mexico State University

David Walker
Educational Tech, Rsrch, & Assessment
Northern Illinois University

J. J. Wang
Dept. of Advanced Educational Studies
California State University, Bakersfield

Dongfeng Wu
Dept. of Mathematics & Statistics
Mississippi State University

Chengjie Xiong
Division of Biostatistics
Washington University in St. Louis

Andrei Yakovlev
Biostatistics and Computational Biology
University of Rochester

Heping Zhang
Dept. of Epidemiology & Public Health
Yale University

INTERNATIONAL

Mohammed Ageel
Dept. of Mathematics, & Graduate School
King Khalid University, Saudi Arabia

Mohammad Fraiwan Al-Saleh
Department of Statistics
Yarmouk University, Irbid-Jordan

Keumhee Chough (K.C.) Carriere
Mathematical & Statistical Sciences
University of Alberta, Canada

Michael B. C. Khoo
Mathematical Sciences
Universiti Sains, Malaysia

Debasis Kundu
Department of Mathematics
Indian Institute of Technology, India

Christos Koukouvinos
Department of Mathematics
National Technical University, Greece

Lisa M. Lix
Dept. of Community Health Sciences
University of Manitoba, Canada

Takis Papaioannou
Statistics and Insurance Science
University of Piraeus, Greece

Nasrollah Saebi
Computing, Information Systems & Math
Kingston University, UK

Keming Yu
Department of Statistics
University of Plymouth, UK

Journal Of Modern Applied Statistical Methods

Invited Articles

2 – 10	James Algina,
H. J. Keselman,
Randall D. Penfield	Confidence Intervals for the Squared
Multiple Semipartial Correlation Coefficient
11 – 20	Rand R. Wilcox	On a Test of Independence via Quantiles that is
Sensitive to Curvature

Regular Articles

21 – 38	D. Roland Thomas,
PengCheng Zhu,
Bruno D. Zumbo,
Shantanu Dutta	On Measuring the Relative Importance of
Explanatory Variables in a Logistic Regression
39 – 55	Gregory R. Hancock,
Michelle M. Buehl	Second-Order Latent Growth Models with
Shifting Indicators
56 – 76	Guili Zhang,
James Algina	Coverage Performance of the Non-Central F -based
and Percentile Bootstrap Confidence Intervals for
Root Mean Square Standardized Effect Size in
One-Way Fixed-Effects ANOVA
77 – 93	Tim Moses,
Alan Klockars	An Evaluation of Standard, Alternative, and
Robust Slope Test Strategies
94 – 100	Weiming Ke,
Rui Yao	Selection of Non-Regular Fractional Factorial
Designs When Some Two-Factor Interactions
are Important
101 – 119	Hua Fang,
Gordon P. Brooks,
Maria L. Rizzo,
Kimberly A. Espy,
Robert S. Barcikowski	A Monte Carlo Power Analysis of Traditional
Repeated Measures and Hierarchical Multivariate
Linear Models in Longitudinal Data Analysis
120 – 139	Massimiliano Pastore,
Massimo Nucci,
Giovanni Galfano	Comparing Different Methods for Multiple
Testing in Reaction Time Data
140 – 151	Zawar Hussain,
Javid Shabbir	Logit Estimation Using Warner's Randomized
Response Model
152 – 157	David A. Walker	Estimating How Many Observations are
Needed to Obtain a Required Level of Reliability

158 – 179	S. Jonathan Mends-Cole	Probability of Coverage and Interval Length for Two-Group Techniques Assessing the Median and Trimmed Mean
180 – 186	Stephen Sharp	Tests for Independence in Two-Way Contingency Tables with Small Samples
187 – 197	Shou Hsing Shih, Chris P. Tsokos	A Weighted Moving Average Process for Forecasting
198 – 204	Madhusudan Bhandary	Estimation of Covariance Matrix in Signal Processing When the Noise Covariance Matrix is Arbitrary
205 – 222	Moonseong Heo	Utility of Weights for Weighted Kappa as a Measure of Interrater Agreement on Ordinal Scale
223 – 233	W. Holmes Finch, Brian F. French	Using Exploratory Factor Analysis for Locating Invariant Referents in Factor Invariance Studies
234 – 245	Mark J. Gierl, Ying Cui, Steve Hunka	Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests
246 – 252	Kayode Ayinde, Ladoke Akintol, J. O. Olaomi	Robustness of Some Estimators of Linear Model with Autocorrelated Error Terms When Stochastic Regressors are Normally Distributed
253 – 258	Meral Cetin, Serpil Aktas	Confidence Intervals Based on Robust Estimators
259 – 274	Haiyan Chen, Howard H. Stratton	Test for Spatio-Temporal Counts Being Poisson
275 – 285	Matthew E. Elam, Kenneth E. Case	Two-Stage Short-Run (X, MR) Control Charts
286 – 296	Chengjie Xiong, Feng Gao, Yan Yan, Jingqin Luo, YunJu Sung, Gang Shi	Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease

297 – 303 **Dongfeng Wu,**
 Ricolindo L. Cariño,
 Xiaoqin Wu When Sensitivity is a Function of Age and Time
Spent in the Preclinical State in Periodic
Cancer Screening

304 – 313 **Krishnan Rajendran,**
 T. Ramamurthy,
 Sujit K. Bhattacharya Log-linear Model to Assess Socioeconomic and
Environmental Factors with Childhood Diarrhea
Using Hospital Based Surveillance

Brief Reports

314 – 317 **W. J. Hurley** On the Length of NHL Shootouts

Statistical Software Applications & Review

318 – 330 **Kimberly Crimin,**
 Asheber Abebe,
 Joseph W. McKean Robust General Linear Models and Graphics via a
User Interface (Web RGLM)

331 – 345 **Roberto M.-Espiñeira,**
 Joseph M. Hilbe Effect on Recreation Benefit Estimates from
Correcting for On-Site Sampling Biases and
Heterogeneous Trip Overdispersion in Count Data
Recreation Demand Models (STATA)

346 – 357 **Michele Scagliarini,**
 Raffaele Vermiglio Computing Multivariate Process Capability
Indices (Excel)

Translations, Ephemerals, & Biographies

358 – 367 **Daniel S. Drucker,**
 Claude Schochet,
 John Cuzzocrea,
 Shlomo S. Sawilowsky Jacques Salomon Hadamard and the Use of
Symbols in Teaching Differential Calculus

JMASM is an independent print and electronic journal (<http://tbf.coe.wayne.edu/jmasm>), publishing (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.

Editorial Assistants: **Jonathan Lent, Michael Lance, Laura Rolands**

Internet Sponsor: **Paula C. Wood**, Dean, College of Education, Wayne State University

Cushing-Malloy, Inc. Internet: www.cushing-malloy.com	(888) 295-7244 toll-free (Phone) (734) 663-5731 (Fax)	Sales & Information: skehoe@cushing-malloy.com
--	--	--

INVITED ARTICLES

Confidence Intervals for the Squared Multiple Semipartial Correlation Coefficient

James Algina
University of Florida

H. J. Keselman
University of Manitoba

Randall D. Penfield
University of Miami

The squared multiple semipartial correlation coefficient is the increase in the squared multiple correlation coefficient that occurs when two or more predictors are added to a multiple regression model. Coverage probability was investigated for two variations of each of three methods for setting confidence intervals for the population squared multiple semipartial correlation coefficient. Results indicated that the procedure that provides coverage probability in the $[.925, .975]$ interval for a 95% confidence interval depends primarily on the number of added predictors. Guidelines for selecting a procedure are presented.

Key words: Squared multiple semipartial correlation; effect size; asymptotic and bootstrap confidence intervals.

Introduction

A commonly used effect size (ES) in multiple regression analysis is the increase in R^2 when

James Algina (algina@ufl.edu) is a Professor of Research and Evaluation Methodology. His research interests are in applied statistics and psychometrics. Email him at algina@ufl.edu. H. J. Keselman is a Professor of Psychology. His research interests are in applied statistics. Email him at kesel@ms.umanitoba.ca. Randall D. Penfield is an Associate Professor of Education. His research interests are in educational measurement and psychometrics. Email him at penfield@miami.edu.

one independent variable (X_j) is added to the model. This ES, which is called the squared semipartial correlation coefficient, and is often symbolized by ΔR^2 , measures the strength of relationship between X_j and the dependent variable (Y), controlling for the other independent variables in the model. This coefficient ΔR^2 can also be used when several variables are added to the model. In this context, ΔR^2 is called the squared multiple semipartial correlation coefficient (Pedhazur, 1997) or the semipartial R^2 (Cohen, Cohen, West, & Aiken, 2003) and measures the strength of association between Y and the added independent

variables, controlling for the other independent variables in the model.

Hedges and Olkin (1981) presented methods for calculating the asymptotic sampling covariance matrix for commonality components (See also Mood, 1969, 1971). These results can be used to construct a confidence interval (CI) for $\Delta\rho^2$, the population ES estimated by ΔR^2 . Olkin and Finn (1995) presented a method equivalent to the Hedges and Olkin method, and illustrated the new method for the case in which there is one independent variable in the model in addition to X_j (i.e., for the case of two independent variables). Alf and Graf (1999) simplified the method and showed how to apply it in the general case of p predictors, and Graf and Alf (1999) developed a computer program that computes the CI. Algina and Moulder (2001) found that when the squared semipartial correlation coefficient is of interest, researchers would need *very* large samples sizes (n) to achieve adequate coverage probability for $\Delta\rho^2$. Algina, Keselman, and Penfield (2007) found that it was possible to obtain much better coverage probability, with smaller sample sizes, if percentile bootstrapping methods were used for setting CIs for the squared semipartial correlation coefficient, rather than relying on the asymptotic intervals. The purpose of the present paper was to investigate whether asymptotic or percentile bootstrap intervals would result in adequate coverage probability for $\Delta\rho^2$ when a squared multiple semipartial correlation coefficient is of interest.

Method

Coverage probability was estimated for the asymptotic and two percentile bootstrap CIs. Specifically, simulation was used to estimate coverage probability for combinations of p , k , n , ρ_f^2 , and ρ_r^2 , where ρ_f^2 is the population squared multiple correlation coefficient for a model with p predictors (the full model), and ρ_r^2 is the population squared multiple correlation coefficient for a model with k predictors (the reduced model) that are a proper subset of the p predictors. In the conditions we planned to

investigate, the number of predictors in the full model ranged from $p=3$ to $p=9$ in steps of 2. After reviewing the results we added conditions with $p=8$ predictors in the full model. The difference in the number of predictors in the full and reduced models (i.e., $p-k$) ranged from 2 to $p-1$ in steps of 1. The squared multiple correlation coefficient for the predictors in the reduced model ranged from .00 to .50 in steps of .10. The squared multiple correlation coefficient for the full model ranged from ρ_r^2 to $\rho_r^2 + .10$ in steps of .01 and from $\rho_r^2 + .10$ to $\rho_r^2 + .20$ in steps of .05. Sample size ranged from 50 to 200 in steps of 50. The predictors and the dependent variable were distributed as a multivariate normal distribution.

Each of the 7176 combinations of p , k , n , ρ_f^2 , and ρ_r^2 was replicated 5000 times in order to estimate coverage probability. For each replication, six 95% CIs were constructed: two CIs was constructed by using two variations of (a) the asymptotic method; (b) bootstrapping $\Delta R^2 = R_f^2 - R_r^2$; and (c) bootstrapping $\Delta R_c^2 = R_{f,c}^2 - R_{r,c}^2$, the difference in corrected values of R^2 , where

$$R_{f,c}^2 = R_f^2 - \frac{p}{n-p-1}(1-R_f^2),$$

and

$$R_{r,c}^2 = R_r^2 - \frac{k}{n-k-1}(1-R_r^2).$$

For each CI, the proportion of the replications that contained $\Delta\rho^2 = \rho_f^2 - \rho_r^2$ estimated the probability of coverage.

Under multivariate normality of the predictors and criterion, the asymptotic variance of R_f^2 is

$$\psi_f^2 = \frac{4\rho_f^2(1-\rho_f^2)^2}{n} \quad (1)$$

(Stuart, Ord, & Arnold, 1999). The asymptotic variance of R_r^2 is obtained by substituting r for f in the subscripting. According to Alf and Graf

C. I. FOR SQUARED MULTIPLE SEMIPARTIAL CORRELATION

(1999), the asymptotic covariance between R_r^2 and R_f^2 is

$$\psi_{fr} = \frac{4\rho_f\rho_r[.5(2\rho_r/\rho_f - \rho_f\rho_r)]}{n} \times \frac{(1 - \rho_f^2 - \rho_r^2 - \rho_r^2/\rho_f^2) + \rho_r^3/\rho_f^3}{n}. \quad (2)$$

The asymptotic variance of ΔR^2 is

$$\sigma_\infty^2 = \psi_f^2 + \psi_r^2 - 2\psi_{fr},$$

and an asymptotically correct $100(1-\alpha)\%$ CI for $\Delta\rho^2$ is

$$\Delta R^2 \pm z_{\alpha/2}\sigma_\infty,$$

where $z_{\alpha/2}$ is a z critical value (Alf & Graf, 1999). In practice, the asymptotic variance is estimated by substituting R_f^2 and R_r^2 for ρ_f^2 and ρ_r^2 , respectively.

Initial results indicated that in some conditions in which $\rho_f^2 = \rho_r^2$, the asymptotic CI resulted in coverage probabilities above .99. To address this problem, the lower limit of the asymptotic CI was modified. Specifically, if the lower limit of the traditional asymptotic CI was less than or equal to zero, but the F test of $H_0:\Delta\rho^2=0$ was significant, the lower limit was set to a small value larger than zero. In our simulations the lower limit was set equal to .001.

To apply the traditional percentile bootstrap, as described in Wilcox (2003), to ΔR^2 the following steps were completed, with the first two steps completed B times.

1. A sample of size n was randomly selected with replacement from the simulated participants.
2. ΔR^2 was calculated for the sample drawn in step1.
3. Once the B values of ΔR^2 were obtained, they were ranked from low to high. The lower limit of the $100(1-\alpha)\%$ CI for $\Delta\rho^2$ was determined by finding the

$\lceil \langle B(\alpha/2) \rangle + 1 \rceil^{\text{th}}$ estimate in the rank order, where $\langle B(\alpha/2) \rangle$ indicates rounding $B(\alpha/2)$ to the nearest whole number; the upper limit was determined by finding the $\lceil B - \langle B(\alpha/2) \rangle \rceil^{\text{th}}$ estimate in the rank order.

In all conditions $B=1000$; thus, the bootstrap lower limit was the 26th value and the upper limit was the 975th value in the traditional percentile bootstrap.

A problem with applying the traditional percentile bootstrap to ΔR^2 will occur when $\Delta\rho^2=0$, that is when $\rho_f^2 = \rho_r^2$. Because R_f^2 will infrequently equal R_r^2 , the CI will almost never contain 0 and the estimated coverage probability will be near zero. To address this problem, we modified the traditional percentile bootstrap by incorporating the F test of $H_0:\Delta\rho^2=0$. When the F test was not significant, the lower limit of the CI was set equal to zero; otherwise, the lower limit was determined by using the traditional percentile bootstrap.

To apply the traditional percentile bootstrap to ΔR_c^2 , steps 1 to 3 were applied with ΔR_c^2 replacing ΔR^2 . Initial results indicated that in some conditions in which $\rho_f^2 = \rho_r^2$, this procedure resulted in coverage probabilities above .99. To address this problem, the lower limit of the bootstrap CI was modified. Specifically, if the lower limit of the traditional CI was less than or equal to zero, but the F test of $H_0:\Delta\rho^2=0$ was significant, the lower limit was set to a small value larger than zero. In our simulations the lower limit was set equal to .001. A Visual Basic 6.0 program that computes the traditional and modified percentile bootstrap CIs for $\Delta\rho^2$ is available at <http://plaza.ufl.edu/algina/index.programs.html>

The multiple regression model is $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$. There is no loss in generality if $\alpha=0$ and/or if the variances of the dependent variable and of the independent variables are set equal to 1.0.

According to Browne (1969, 1975), given any set of predictors that has a squared multiple correlation coefficient of ρ^2 with Y , it is always possible to transform the predictors so that (a) the independent variables are mutually uncorrelated and (b) the regression coefficients are equal to any set of values such that

$$\sum_{j=1}^p \beta_j^2 = \sigma_y^2 \rho^2.$$

Therefore, in the simulations (a) all variables had variance equal to one, (b) the independent variables were mutually uncorrelated, and (c) $\beta_1 = \dots = \beta_{k-1} = 0$, $\beta_k = \rho_r$, $\beta_{k+1} = \dots = \beta_{p-1} = 0$, and $\beta_p = \sqrt{\rho_f^2 - \rho_r^2}$.

Thus, the squared multiple correlation coefficient was ρ_r^2 for variables X_1 to X_k and ρ_f^2 for variables X_1 to X_p .

The data were simulated by using the following steps.

1. Generate an $n \times p$ matrix of random variables. Each of the p variables was normally distributed with mean zero and standard deviation one. All np scores were generated to be statistically independent. This matrix is $\mathbf{X}$, the matrix of scores on the independent variables.
2. Generate an $n \times 1$ vector of normally distributed random variables with mean zero and standard deviation one. All n scores were generated to be statistically independent and to be independent of the scores in $\mathbf{X}$. Multiply the generated vector by $\sqrt{1 - \rho_f^2}$. The resulting vector is $\boldsymbol{\varepsilon}$, the vector of residuals.
3. Construct the $p \times 1$ vector $\boldsymbol{\beta}$ in which elements 1 to $k - 1$ are zero and the next element is ρ_r , the elements $k + 1$ to $p - 1$ are zero, and the last element is $\sqrt{\rho_f^2 - \rho_r^2}$.
4. Calculate $n \times 1$ scores on the dependent variable by using $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.

Results

The traditional bootstrap CI using ΔR^2 was modified by setting the lower limit to zero if the F test of $H_0: \Delta \rho^2 = 0$ was not significant; otherwise, the lower limit was determined by using the traditional percentile bootstrap. Although the modification was designed to improve performance when $\Delta \rho^2$ was zero, the modification could affect coverage probability of the modified CI when $\Delta \rho^2$ was zero or larger. Thus, it was important to determine if coverage probability of the bootstrap CI using ΔR^2 was affected by the modification when $\Delta \rho^2$ was not equal to zero. To do this, we focused on the conditions in which $\Delta \rho^2 > 0$. When $\Delta \rho^2 > 0$, for each combination of p , $p - k$, and n , there are 72 combinations of ρ_r^2 and ρ_f^2 . For each combination we tabulated the number of times that the estimated coverage probability was in the interval $[.925, .975]$ for the traditional and modified versions of the bootstrap CI using ΔR^2 . Results indicated that the modification did not reduce the number of conditions in which the interval contained the estimated coverage probability. The asymptotic CI and traditional bootstrap CI using ΔR_c^2 were modified by changing the lower limit to .001 when the traditional lower limit was less or equal to zero and the F test of $H_0: \Delta \rho^2 = 0$ was significant. For our values of ρ_r^2 and ρ_f^2 , this modification could only affect the performance of these CIs when $\Delta \rho^2$ was zero (and in general could only affect the CI if $\Delta \rho^2 < .001$). Therefore, the following results describe performance of the modified CIs.

For each combination of p , $p - k$, and n , we tabulated the number of times out of 78 possible combinations of ρ_r^2 and ρ_f^2 in which the estimated coverage probability was in the interval $[.925, .975]$ for the modified version of the CIs. The results are presented in Table 1. For each combination, the bold value indicates the

C. I. FOR SQUARED MULTIPLE SEMIPARTIAL CORRELATION

method(s) that best controlled probability coverage. Ties between methods are indicated by an underlined bolded result. For $p-k=2$, there were combinations of p and n in which the bootstrap CI using ΔR^2 performed as well or better than the other CIs. However, for other values of $p-k$, either the asymptotic or bootstrap CI using ΔR_c^2 performed better than the bootstrap CI using ΔR^2 . The relative performance of the modified asymptotic bootstrap and the modified percentile bootstrap using ΔR_c^2 depended on $p-k$. When $p-k=3$, the modified percentile bootstrap tended to work better if $p \leq 7$; otherwise, the two procedures worked about equally well. The modified percentile bootstrap tended to work better when $p-k$ was between 4 or 5. When $p-k=6$, the two procedures worked about equally well, particularly when the sample size was at least 150. For $p-k$ larger than six, a value that could only occur in our design with 8 or 9 predictors in the full model, the modified asymptotic bootstrap had better control of coverage probability.

Inspection of the results suggests that for $p-k=2$, the relative performance of the modified asymptotic CI and the modified bootstrap CI using ΔR^2 depends on $\Delta \rho^2$. The results also suggest that larger sample sizes are required to achieve control of coverage probability when $\Delta \rho^2$ is small. To illustrate these effects, we tabulated (Table 2) the number of times out of the 18 possible combinations of ρ_r^2 and ρ_f^2 for $\Delta \rho^2 \leq .02$ and the number of times out of the 60 possible combinations of ρ_r^2 and ρ_f^2 for $\Delta \rho^2 \geq .03$ (Table 3), that the estimated coverage probability was in the interval $[.925, .975]$. When $\Delta \rho^2 \leq .02$ and $p-k=2$, neither the modified bootstrap CI using ΔR^2 nor the modified asymptotic CI was consistently more effective than the other over all values of p . Temporarily defining effective as all estimated coverage probabilities within the interval $[.925, .975]$, the modified bootstrap CI using ΔR^2 was effective at a smaller size than

was the asymptotic CI (See Table 3) for $\Delta \rho^2 \geq .03$ and $p-k=2$. In regard to sample size required to achieve good control of coverage probability, the following comments apply to all combinations of p and $p-k$, with the exception of conditions in which there were nine predictors in the full model and no more than two in the reduced model. When $\Delta \rho^2 \leq .02$, the sample size required for at least one of the methods to be effective was 200 in some conditions. When $\Delta \rho^2 \geq .03$, a sample size of 50 to 100 was sufficient for at least one of the methods to be effective.

Discussion

We investigated coverage probability for the asymptotic CI and two percentile bootstrap CIs for $\Delta \rho^2$ in multiple linear regression analyses when predictors and criterion were normally distributed and $\Delta \rho^2$ described the strength of association for several predictors. We also investigated modified versions of these CIs. In general, the modified methods worked at least as well as their unmodified counterparts.

Specifically, results indicated that the traditional and modified bootstrap CI using ΔR^2 performed poorly, except when $p-k=2$. Algina, et al. (2007) reported that when $\Delta \rho^2$ describes the strength of association for one predictor ($p-k=1$), using the modified percentile bootstrap with ΔR^2 to set a CI for $\Delta \rho^2$ resulted in good coverage probability in a wide range of conditions. Thus, the results, for the case in which $p-k=1$, do not generalize to the cases in which $\Delta \rho^2$ describes the strength of association for more than one predictor ($p-k > 2$).

The traditional and modified asymptotic CIs worked well in a variety of conditions. These results are contrary to results reported by Algina and Moulder (2001) who investigated the case of $p-k=1$ and reported that the traditional asymptotic CI tended to work poorly in many conditions with sample sizes of 200 or

Table 1. Number of Coverage Probability Estimates (out of 78) Inside the $[\text{.925}, \text{.975}]$ Interval for the Modified CIs

		Modified Bootstrap on ΔR^2				Modified Asymptotic				Modified Bootstrap on ΔR_c^2			
		n				n				n			
p	$p-k$	50	100	150	200	50	100	150	200	50	100	150	200
3	2	55	<u>71</u>	<u>76</u>	75	43	<u>71</u>	<u>76</u>	78	59	61	69	69
5	2	66	73	76	77	26	61	74	78	47	63	67	69
	3	20	36	48	55	66	72	74	77	78	78	77	78
	4	9	11	13	22	70	71	72	71	78	78	78	78
7	2	68	73	76	<u>78</u>	23	35	70	<u>78</u>	33	60	67	68
	3	37	46	56	60	64	72	74	78	77	76	76	77
	4	13	16	25	28	73	71	72	72	78	78	78	78
	5	9	8	10	12	54	75	77	76	77	78	78	78
	6	10	6	8	9	29	73	<u>78</u>	<u>78</u>	18	70	<u>78</u>	<u>78</u>
8	2	70	75	77	78	23	24	67	74	30	59	65	69
	3	41	51	57	60	60	72	75	77	67	74	74	76
	4	20	21	28	34	71	73	72	72	78	78	78	78
	5	13	10	13	15	53	78	76	76	77	78	78	78
	6	10	6	7	11	25	72	77	<u>78</u>	23	73	78	<u>78</u>
	7	10	6	6	7	12	42	76	78	4	26	54	64
9	2	62	75	77	78	17	20	57	74	28	58	66	69
	3	45	52	58	62	63	69	75	78	56	73	72	77
	4	23	25	29	38	71	72	72	72	78	78	78	78
	5	13	12	15	19	51	77	77	77	77	78	78	78
	6	10	8	9	11	28	74	<u>78</u>	<u>78</u>	37	73	<u>78</u>	<u>78</u>
	7	12	6	7	9	24	45	66	69	23	41	55	59
	8	10	6	6	6	5	19	38	55	1	7	27	38

C. I. FOR SQUARED MULTIPLE SEMIPARTIAL CORRELATION

Table 2. Number of Coverage Probability Estimates (out of 18) Inside the $[\text{.925}, \text{.975}]$ Interval for Modified CIs and $\Delta\rho^2 \leq .02$

		Modified Bootstrap on ΔR^2				Modified Asymptotic				Modified Bootstrap on ΔR_c^2			
		n				n				n			
p	$p-k$	50	100	150	200	50	100	150	200	50	100	150	200
3	2	6	11	16	15	12	14	17	18	7	7	10	9
5	2	8	13	16	17	12	16	17	18	6	7	8	9
	3	6	6	6	6	9	12	14	17	18	18	17	18
	4	8	6	6	6	15	11	12	11	18	18	18	18
7	2	8	13	<u>16</u>	<u>18</u>	12	14	<u>16</u>	<u>18</u>	6	6	8	8
	3	8	6	6	7	13	12	14	18	18	16	16	17
	4	7	6	6	6	17	12	12	12	18	18	18	18
	5	8	6	6	6	15	17	17	16	17	18	18	18
	6	10	6	6	6	5	18	<u>18</u>	<u>18</u>	0	11	<u>18</u>	<u>18</u>
8	2	11	<u>15</u>	<u>17</u>	18	12	<u>15</u>	<u>17</u>	17	6	6	7	9
	3	7	6	6	6	13	12	15	17	17	15	14	16
	4	9	6	6	6	17	13	12	12	18	18	18	18
	5	10	6	6	6	13	18	16	16	17	18	18	18
	6	10	6	6	6	4	18	<u>18</u>	<u>18</u>	0	13	<u>18</u>	<u>18</u>
	7	10	6	6	6	0	10	18	18	0	0	1	4
9	2	15	<u>15</u>	17	<u>18</u>	12	<u>15</u>	15	<u>18</u>	6	6	8	9
	3	7	6	6	7	13	<u>13</u>	15	18	16	<u>13</u>	12	17
	4	8	6	6	6	17	12	12	12	18	18	18	18
	5	9	6	6	6	13	17	17	17	17	18	18	18
	6	9	6	6	6	4	18	<u>18</u>	<u>18</u>	0	13	<u>18</u>	<u>18</u>
	7	12	6	6	6	0	8	17	18	0	0	2	6
	8	10	6	6	6	0	1	4	8	0	0	0	0

Table 3. Number of Coverage Probability Estimates (out of 60) Inside the $[.925, .975]$ Interval for Modified CIs and $\Delta\rho^2 \geq .03$

		Modified Bootstrap on $ \Delta R^2 $				Modified Asymptotic				Modified Bootstrap on $ \Delta R_c^2 $			
		$ n $				$ n $				$ n $			
$ p $	$ p - k $	50	100	150	200	50	100	150	200	50	100	150	200
3	2	49	60	60	<u>60</u>	31	57	59	<u>60</u>	52	54	59	<u>60</u>
5	2	58	60	60	<u>60</u>	14	45	57	<u>60</u>	41	56	59	<u>60</u>
	3	14	30	42	49	57	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
	4	1	5	7	16	55	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
7	2	60	60	60	<u>60</u>	11	21	54	<u>60</u>	27	54	59	<u>60</u>
	3	29	40	50	53	51	<u>60</u>	<u>60</u>	<u>60</u>	59	<u>60</u>	<u>60</u>	<u>60</u>
	4	6	10	19	22	56	59	<u>60</u>	<u>60</u>	60	60	<u>60</u>	<u>60</u>
	5	1	2	4	6	39	58	<u>60</u>	<u>60</u>	60	60	<u>60</u>	<u>60</u>
	6	0	0	2	3	24	55	<u>60</u>	<u>60</u>	18	59	<u>60</u>	<u>60</u>
8	2	59	60	60	<u>60</u>	11	9	50	57	24	53	58	<u>60</u>
	3	34	45	51	54	47	60	<u>60</u>	<u>60</u>	50	59	<u>60</u>	<u>60</u>
	4	11	15	22	28	54	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
	5	3	4	7	9	40	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
	6	0	0	1	5	21	54	59	<u>60</u>	23	60	60	<u>60</u>
	7	0	0	0	1	12	32	58	<u>60</u>	4	26	53	<u>60</u>
9	2	47	60	60	<u>60</u>	5	5	42	56	22	52	58	<u>60</u>
	3	38	46	52	55	50	56	<u>60</u>	<u>60</u>	40	60	<u>60</u>	<u>60</u>
	4	15	19	23	32	54	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
	5	4	6	9	13	38	<u>60</u>	<u>60</u>	<u>60</u>	60	<u>60</u>	<u>60</u>	<u>60</u>
	6	1	2	3	5	24	56	<u>60</u>	<u>60</u>	37	60	<u>60</u>	<u>60</u>
	7	0	0	1	3	24	37	49	51	23	41	53	53
	8	0	0	0	0	5	18	34	47	1	7	27	38

C. I. FOR SQUARED MULTIPLE SEMIPARTIAL CORRELATION

less. Algina and Moulder did not report results on the modified asymptotic CI, but the modification used in this paper is designed to improve performance when $\Delta\rho^2$ is zero, so that it is unlikely that using the modification with $p-k=1$ would overcome the problems that Algina and Moulder reported.

Unfortunately, although the modified asymptotic CI and the modified bootstrap CI on ΔR_c^2 worked better than the competitors investigated in this study, neither of these CIs worked well in all of the conditions we investigated. Particularly problematic was the condition in which the number of predictors in the full model was nine and the number of predictors in the reduced model was no larger than two. Defining adequate control of coverage probability by at least 77 of the 78 combinations of ρ_r^2 and ρ_f^2 for each combination of p and predictors in the reduced model was no larger than two. Defining adequate control of coverage probability by at least 77 of the 78 combinations of ρ_r^2 and ρ_f^2 for each combination of p and $p-k$, we offer the following recommendations for a CI method and sample size in order to achieve adequate control of coverage probability.:

(a) If $p-k=2$ and $p \leq 7$, the modified asymptotic CI should be used with a sample size of at least 200. For $p \geq 8$, a sample size of 150 and the modified bootstrap CI using ΔR^2 should be used;

(b) If $p-k=3$ and $p \leq 7$, the modified bootstrap CI using ΔR_c^2 should be used with a sample size of at least 50. If $p \geq 8$ the modified asymptotic CI should be used with a sample size of at least 200;

(c) If $p-k=4$ or 5 and $p \leq 9$, the modified bootstrap CI using ΔR_c^2 should be used with a sample size of at least 50;

(d) If $p-k=6$ (and therefore $p \geq 7$), the modified bootstrap CI using ΔR_c^2 should be used with a sample size of at least 150;

(e) If $p-k \geq 7$ (and therefore $p \geq 8$), the modified asymptotic CI should be used and

the sample size should 200 if $p=8$ and should be larger than 200 if $p=9$.

References

- Alf, E. F., Jr., & Graf, R. G. (1999). Asymptotic confidence limits for the difference between two squared multiple correlations: A simplified approach. *Psychological Methods*, 4, 70-75.
- Algina, J., Keselman, H. J., & Penfield, R. D. (2007). Confidence intervals for an effect size measure in multiple linear regression. *Educational and Psychological Measurement*, 67, 207-218.
- Algina, J., & Moulder, B. C. (2001). Sample sizes for confidence intervals on the increase in the squared multiple correlation coefficient. *Educational and Psychological Measurement*, 61, 633-649.
- Browne, M. W. (1969). *Precision of prediction* (Research Bulletin 69-69). Princeton, NJ: Educational Testing Service.
- Browne, M. W. (1975). Predictive validity of a linear regression equation. *British Journal of Mathematical & Statistical Psychology*, 28, 79-87.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression and correlation for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.
- Graf, R. G., & Alf, E. F., Jr. (1999). Asymptotic confidence limits for partial and squared partial correlations. *Applied Psychological Measurement*, 23, 116-119.
- Hedges, L. V., & Olkin, I. (1981). The asymptotic distribution of commonality components. *Psychometrika*, 46, 331-336.
- Mood, A. M. (1969). Macro-analysis of the American educational system. *Operations Research*, 17, 770-784.
- Mood, A. M. (1971). Partitioning variance in multiple regression analyses as a tool for developing learning models. *American Educational Research Journal*, 8, 191-202.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Belmont, CA: Wadsworth.
- Olkin, I., & Finn, J. D. (1995). Correlations redux. *Psychological Bulletin*, 118, 155-164.
- Stuart, A., Ord, J. K., & Arnold, S. (1999). *The advanced theory of statistics* (6th ed., Vol. 2A). New York: Oxford University Press.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

On a Test of Independence via Quantiles that is Sensitive to Curvature

Rand R. Wilcox
University of Southern California

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, be a random sample from some $p+1$ variate distribution where $\mathbf{X}_i$ is a vector having length p . Many methods for testing the hypothesis that Y is independent of $\mathbf{X}$ are relatively insensitive to a broad class of departures from independence. Power improvements focus on the median of Y or some other quantile and test the hypothesis that the regression surface is a horizontal plane versus some unknown form. A wild bootstrap method (Stute et al. 1998) can be used based on quantiles, but with small or moderate sample sizes, control over the probability of a Type I error can be unsatisfactory when sampling from asymmetric distributions. He and Zhu (2003) is readily adapted to testing the hypothesis that the conditional γ quantile of Y does not depend on $\mathbf{X}$ where critical values are determined via simulations. A modification is suggested that avoids the need for simulations to obtain critical values, and perform wells in terms of Type I errors even when sampling from asymmetric distributions.

Keywords: Curvature, quantile regression, robust methods.

Introduction

Let $(Y_i, \mathbf{X}_i)$, $i = 1, \dots, n$, be a random sample from some $p+1$ variate distribution where $\mathbf{X}_i$ is a vector of length p . Certainly one of the most common methods for attempting to detect an association between Y and $\mathbf{X}$ is to test the hypothesis that the corresponding (Pearson) correlations are zero using Student's t test. One well-known limitation of this approach is that true associations can be missed due to curvature.

Another concern is that classic methods can be highly unsatisfactory in terms of controlling the probability of a Type I error.

For example, a general method for testing the hypothesis of independence among sets of variables, which assumes multivariate normality, is available (e.g., Muirhead, 1982, chapter 11). As a special case, the method can be used to test

$$H_0 : \rho_{y1} = \dots = \rho_{yp} = 0$$

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. He is the author of eight books, the most recent of which is *Basic Statistics: Understanding Conventional Methods and Modern Insights* (2009, NY: Oxford University Press).

where ρ_{yj} is Pearson's correlation between Y and X_j , $j=1, \dots, p$. But it is known that the level of this test cannot be controlled in an adequate fashion (e.g., Reddon, Jackson, & Schopflocher, 1985; Wilcox, 1997). One could use a method based on Fisher's r -to- z transformation, but this can be unsatisfactory when sampling from

nonnormal distributions because under general conditions, Fisher's transformation is not even asymptotically correct; it results in using the wrong standard error (Duncan and Layard, 1973). Geiser and Randle (1997) derived a nonparametric method, but it assumes that Y has a symmetric distribution and that the distribution of $\mathbf{X}$ is elliptically symmetric, a restriction that is avoided here. There are many robust correlation coefficients as well as robust regression methods that might be used (e.g., Wilcox, 2005), as well as improved methods when focusing on Pearson's correlation (e.g., Boik & Haaland, 2006), but it is evident that these methods can miss true associations that are detected when focusing on quantiles via the method considered here.

Let $m(\mathbf{X}) = E(Y | \mathbf{X})$ and let $\mu_y = E(Y)$. A general and relatively flexible approach to detecting dependence is to test

$$H_0 : m(\mathbf{X}) = \mu_y \quad (1)$$

versus the alternative hypothesis that $m(\mathbf{X})$ depends in some (unspecified) manner on $\mathbf{X}$, possibly in a nonlinear fashion. A test of (1) can be performed using a method that stems from general theoretical results reported by Stute, Gonzalez Manteiga and Presedo Quindimil (1998) who were concerned with testing the hypothesis that a regression surface belongs to a specified family of functions. Unlike conventional methods, by design the method is not sensitive to heteroscedasticity. That is, if we model the data with $Y = m(\mathbf{X}) + \lambda(\mathbf{X})\varepsilon$, where the error term ε independent of $\mathbf{X}$, and $\lambda(\mathbf{X})$ is some unknown function. The assumption $\lambda(\mathbf{X}) \equiv 1$ (homoscedasticity) is not made nor required when testing (1). In principle, the method can be extended by replacing the conditional mean of Y with the median or some other robust estimator. When ε has a symmetric distribution, control over the probability of a Type I error has been found to be satisfactory in simulations, but when ε has an asymmetric distribution, this is no longer the case (Wilcox, 2007).

Let $Y_\gamma(\mathbf{X})$ be the conditional γ quantile of Y given $\mathbf{X}$. A general method derived by He and Zhu (2003) is readily adapted to the problem of testing

$$H_0 : Y_\gamma(\mathbf{X}) = Y_\gamma \quad (2)$$

where Y_γ is the γ quantile of the marginal distribution of Y . The .5 quantile is perhaps the most obvious choice, but in some situations associations are more pronounced when considering other quantiles, and in some cases other quantiles are intrinsically interesting. The He and Zhu method is based in part on using simulations to estimate the null distribution of their test statistic. Execution time is reasonably low with small sample sizes, but despite the speed of modern

computers, execution time can be high. For example, with a sample size of $n=100$ and $p=4$, execution time was over 8 minutes on a SUN BLADE 150.

The goal in this paper is to suggest a simple modification of the method derived by He and Zhu (2003) that, for a wide range of situations, can be used to test (2) without resorting to simulation estimates of critical values.

Simulation results reported here find that the actual level of the test is reasonably close to the nominal level, even when sampling from asymmetric distributions and there is a fair degree of heteroscedasticity.

Method

Let $\mathbf{x}$ be the n by $(p+1)$ matrix with the first column containing all ones and the remaining p columns are the columns of $\mathbf{X}$. Following He and Zhu, it is assumed that the design has been normalized so that $n^{-1} \sum \mathbf{x}_j \mathbf{x}_j' - \mathbf{I} = o(1)$.

Let $r_i = Y_i - \hat{Y}_\gamma$, where $\hat{Y}_\gamma$ is some estimate of the γ th quantile of Y . Here, the focus is on the quartiles. For the .5 quantile, $\hat{Y}_{.5}$ is taken to be the usual sample median. For the lower and upper quartiles, the so-called ideal

fourths are used (e.g., Frigge, Hoaglin, and Iglewicz, 1989), which are computed as follows. Let $j=(n/4)+(5/12)$, rounded down to the nearest integer. Let

$$h = \frac{n}{4} + \frac{5}{12} - j.$$

Then the estimate of the lower quartile is given by

$$q_1 = (1-h)Y_{(j)} + hY_{j+1},$$

where $Y_{(1)} \leq \dots \leq Y_{(n)}$. Letting $k=n-j+1$, the estimate of the upper quartile, is

$$q_2 = (1-h)Y_{(k)} + hY_{k-1}$$

There are many other quantile estimators, comparisons of which are reported by Parrish (1990) as well as Dielman, Lowry, and Pfaffenberger (1994). Perhaps they offer some practical advantage for the situation at hand, but this is not pursued here.

Following He and Zhu, for any $\mathbf{x}$, $\mathbf{t} \in R^p$, $\mathbf{x} \leq \mathbf{t}$ if and only if each component of $\mathbf{x}$ is less than or equal to each component of $\mathbf{t}$. Let $\psi(r) = \gamma I(r > 0) + (\gamma - 1)I(r < 0)$, and let

$$\mathbf{R}_n \mathbf{t} = n^{-1/2} \sum_{j=1}^n \psi(r_j) I(\mathbf{x}_j \leq \mathbf{t})$$

The He and Zhu test statistic, for the situation at hand, is

$$T_n = \max_{\|\mathbf{a}\|=1} n^{-1} \sum (\mathbf{a}' \mathbf{R}_n(\mathbf{x}_j))^2 \quad (3)$$

the largest eigenvalue of $n^{-1} \sum \mathbf{R}_n(\mathbf{x}_i) \mathbf{R}_n'(\mathbf{x}_i)$.

A simple strategy for determining an appropriate critical value is to temporarily assume normality, use simulations to approximate the $1-\alpha$ quantile of the null distribution, say c , and then reject the null hypothesis if $T_n \geq c$ even when sampling from a non-normal distribution. It was found, however, that this strategy performed in an unsatisfactory manner, in simulations, when sampling from heavy-tailed distributions.

(The actual Type I error probability can exceed .08 when testing at the .05 level.) However, a simple modification was found to give better results. The modification consists of using a different partial ordering on the design space; otherwise the test statistic is computed in the same manner as T_n . Let

$$\mathbf{R}_n(\mathbf{x}_i) = n^{-1/2} \sum_{k=1}^n \psi(r_k) \mathbf{x}_k I(\mathbf{x}_k \leq \mathbf{x}_i)$$

For fixed j , let U_{ij} be the ranks of the n values in the j th column of $\mathbf{x}$, $j=2, \dots, q$. Let $F_i = \max U_{ij}$, the maximum being taken over $j=2, \dots, q$. If $\mathbf{x}_k \leq \mathbf{x}_i$, then $F_k \leq F_i$. Let

$$\mathbf{W}_i = n^{-1/2} \sum_{k=1}^n \psi(r_k) \mathbf{x}_k I(F_k \geq F_i)$$

The test statistic used here is D_n , the largest eigenvalue of

$$\mathbf{Z} = n^{-1} \sum \mathbf{W}_i \mathbf{W}_i'$$

Numerical checks on this test statistic indicate that it is invariant when the design space, $\mathbf{X}$, is shifted in location. This is in contrast to a related method for testing the fit of a quantile regression model, currently under investigation, which Xuming He (personal communication) pointed out does not enjoy this property. (Centering the design space eliminates this problem, but here this does not seem to be necessary.)

Note that a major component of the test statistic D_n is invariant under monotone transformations of the covariates; only the ranks of the marginal distributions of $\mathbf{X}$ are needed. However, the test statistic can be affected by monotone transformations because this can alter the $\psi(r_i)$ values. But it was found among the simulations reported later in this paper that typically the $\psi(r_i)$ values are altered by a relatively small amount suggesting a simple approach toward determining an appropriate

TEST OF INDEPENDENCE VIA QUANTILES SENSITIVE TO CURVATURE

critical value: Momentarily assume that both $\mathbf{X}$ and ε have standard normal distributions, then use simulations to determine a critical value given n , p and γ , and use this critical value for the more general case where $\mathbf{X}$ and ε do not have normal distributions. (In essence, this is the same strategy used by Gosset to derive Student's T test.)

Simulations are not needed once a critical value has been estimated. (For $p > 1$, all indications are that it suffices to assume that the correlations among the covariates are zero when determining critical values.)

Some Special Cases

Simulations were used to approximate critical values in the manner just described for $p=1, \dots, 8$ predictors; $n=10, 20, 30, 50, 100, 200$ and 400 ; $\gamma=.5, .25$ and $.75$; and $\alpha=.1, .05, .025$ and $.01$. The results are reported in Tables 1-4. Regarding sample sizes not tabled, it was found that there is an approximately linear association between the α level critical value, c_α , and $1/n$ suggesting that a single regression line might be used to determine c_α given $1/n$. However, slightly better control over the probability of a Type I error is obtained by using the critical values in Tables 1-4 and interpolating on $1/n$ for critical values not tabled.

Results

Simulations were used to check the small sample properties of the method just described. Included were situations where $p=1$ and 4 , $\gamma=.5, .25$ and $.75$, and where for $p=4$ there is a common correlation ρ or $.5$. Here the results for $n=20$, $\rho=.5$ and $\gamma=.75$ are reported because the largest deviations from the nominal Type I error probability occurred for this special case. In the simulations, observations were generated with the model

$$Y_i = \lambda(X_{i1})\varepsilon_i,$$

where λ is some function for modeling heteroscedasticity. The distribution of ε was taken to be one of four g-and-h distributions (Hoaglin, 1985), which contains the standard

normal distribution as a special case. If Z has a standard normal distribution, then

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

if $g > 0$, has a g-and-h distribution where g and h are parameters that determine the first four moments. When $g=0$, then

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

The four distributions used here were the standard normal ($g=h=0$), a symmetric heavy-tailed distribution ($h=.2, g=0$), an asymmetric distribution with relatively light tails ($h=0, g=.2$), and an asymmetric distribution with heavy tails ($g=h=.2$).

Table 5 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution considered.

When $g > 0$ and $h > 1/k$, $E(W^k)$ is not defined and the corresponding entry in Table 1 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

The function λ was chosen to reflect three types of variance patterns: $\lambda(\mathbf{X})=1$ (homoscedasticity) $\lambda(\mathbf{X})=X_1^2$, and $\lambda(\mathbf{X})=1+1/(|X_1|+1)$. For convenience, these three λ functions will be called variance patterns VP1, VP2, and VP3.

Each replication in the simulations consisted of generating n vectors for $\mathbf{X}$, n values for ε , determining Y according to equation (3), then applying the test of (2). Here, 1,000 replications were used to estimate the actual probability of a type I error. With 1,000 replications, if the actual probability of a type I error is .05, the standard error associated with the proportion of rejections is .007.

Table 6 shows the estimated Type I error probabilities for $n=20$, $p=4$, a common correlation $\rho=.5$, and $\alpha=.05$. As can be seen, the estimates range between .039 and .071. There are only two situations where the estimate is greater than or equal to .07.

RAND R. WILCOX

Table 1: Critical values for $\gamma=.5$, $1 \leq p \leq 4$

p=1

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.033939	0.04408	0.050923	0.064173
20	0.015323	0.021007	0.027687	0.032785
30	0.010648	0.014778	0.01825	0.023639
50	0.006619	0.009078	0.011691	0.014543
100	0.003156	0.004375	0.005519	0.007213
200	0.001545	0.002232	0.002748	0.003726
400	0.000772	0.001022	0.001371	0.001818

p=2

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.052848	0.061919	0.071347	0.079163
20	0.021103	0.027198	0.031926	0.035084
30	0.013721	0.018454	0.022177	0.026052
50	0.00839	0.01059	0.012169	0.015346
100	0.004262	0.005514	0.007132	0.008417
200	0.001895	0.002416	0.003085	0.003925
400	0.001045	0.001348	0.001579	0.001864

p=3

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.071556	0.082938	0.089555	0.097538
20	0.031061	0.035799	0.043863	0.053712
30	0.019504	0.023776	0.02718	0.030991
50	0.01103	0.013419	0.015557	0.01798
100	0.005634	0.006805	0.007878	0.008808
200	0.002552	0.003604	0.004276	0.005022
400	0.001251	0.001532	0.001801	0.002038

p=4

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.093268	0.101584	0.108734	0.11834
20	0.038678	0.04552	0.051403	0.060097
30	0.024205	0.02936	0.034267	0.039381
50	0.013739	0.015856	0.018066	0.019956
100	0.006468	0.007781	0.009038	0.010127
200	0.003197	0.003934	0.004657	0.005929
400	0.001653	0.001926	0.002364	0.002657

TEST OF INDEPENDENCE VIA QUANTILES SENSITIVE TO CURVATURE

Table 2: Critical values for $\gamma=.5$, $5 \leq p \leq 8$

p=5

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.117217	0.124714	0.129459	0.136456
20	0.048839	0.055609	0.06058	0.067944
30	0.030595	0.035004	0.040434	0.047649
50	0.01694	0.019527	0.022047	0.025313
100	0.008053	0.009779	0.01149	0.013384
200	0.003761	0.004376	0.005098	0.005866
400	0.001895	0.002254	0.002612	0.002939

p=6

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.136962	0.14412	0.149004	0.152667
20	0.055909	0.062627	0.069978	0.08119
30	0.034635	0.040741	0.044161	0.047722
50	0.020165	0.023075	0.025881	0.02848
100	0.009436	0.011247	0.013221	0.015101
200	0.004645	0.005334	0.006041	0.007237
400	0.002278	0.002636	0.002997	0.003669

p=7

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.15618	0.16322	0.17175	0.17714
20	0.07011	0.07705	0.08272	0.09041
30	0.04177	0.04737	0.0531	0.05767
50	0.02338	0.02601	0.0296	0.03261
100	0.01085	0.01256	0.01374	0.01625
200	0.00516	0.00613	0.00686	0.00835
400	0.00253	0.00304	0.00362	0.00397

p=8

n	$\alpha = .100$	$\alpha = .050$	$\alpha = .025$	$\alpha = .001$
10	0.17839	0.18	0.19379	0.19958
20	0.07803	0.08562	0.09151	0.10249
30	0.04599	0.05218	0.05736	0.06263
50	0.02589	0.02973	0.03374	0.03787
100	0.01219	0.01365	0.01548	0.01771
200	0.00589	0.00687	0.00789	0.00852
400	0.00283	0.00324	0.00373	0.00412

RAND R. WILCOX

Table 3: Critical values for $\gamma=.25$ or $.75$, $1 \leq p \leq 4$

p=1

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.029933	0.039598	0.054088	0.062961
20	0.011122	0.014989	0.018154	0.022685
30	0.009207	0.011302	0.014872	0.019931
50	0.004824	0.00704	0.010357	0.013177
100	0.00237	0.003315	0.004428	0.005123
200	0.001106	0.001611	0.001984	0.00265
400	0.000517	0.00068	0.000869	0.001202

p=2

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.044842	0.06026	0.066001	0.087041
20	0.017341	0.022471	0.027371	0.033436
30	0.012121	0.015041	0.018939	0.022644
50	0.006489	0.008461	0.0107	0.013232
100	0.002973	0.004064	0.004911	0.005769
200	0.001515	0.002058	0.002583	0.003114
400	0.000798	0.000993	0.001183	0.001399

p=3

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.063653	0.072975	0.083841	0.097222
20	0.021659	0.027437	0.031875	0.03683
30	0.01529	0.018964	0.021729	0.02896
50	0.008357	0.010072	0.012713	0.015255
100	0.003903	0.004764	0.005577	0.00666
200	0.001914	0.002343	0.002834	0.003465
400	0.00096	0.001147	0.001356	0.001548

p=4

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.085071	0.095948	0.104197	0.11845
20	0.029503	0.034199	0.039543	0.045044
30	0.019203	0.022769	0.026887	0.033482
50	0.01144	0.013555	0.016139	0.018298
100	0.004863	0.005756	0.007385	0.009115
200	0.002635	0.003111	0.003769	0.004216
400	0.001189	0.001435	0.001728	0.001956

TEST OF INDEPENDENCE VIA QUANTILES SENSITIVE TO CURVATURE

Table 4: Critical values for $\gamma=.25$ and $.75$, $5 \leq p \leq 8$

p=5

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.102894	0.114259	0.122545	0.130222
20	0.036733	0.042505	0.048664	0.055457
30	0.024193	0.028806	0.032924	0.03821
50	0.012663	0.014635	0.017276	0.019736
100	0.006106	0.007311	0.00896	0.009745
200	0.003067	0.003615	0.003998	0.004812
400	0.001441	0.001733	0.002079	0.002308

p=6

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.117643	0.126566	0.133107	0.14228
20	0.044309	0.049732	0.053913	0.060513
30	0.028607	0.033826	0.038616	0.043547
50	0.015445	0.017557	0.020041	0.022748
100	0.007335	0.008406	0.009392	0.01092
200	0.003352	0.003815	0.004381	0.005252
400	0.001704	0.002002	0.002339	0.002773

p=7

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.106573	0.113059	0.117388	0.121287
20	0.05217	0.058363	0.064734	0.069749
30	0.030697	0.035507	0.039266	0.044438
50	0.016737	0.019606	0.021254	0.022923
100	0.007767	0.009232	0.010341	0.011471
200	0.003998	0.00459	0.005507	0.006217
400	0.001903	0.002175	0.002519	0.002859

p=8

n	$\alpha=.100$	$\alpha=.050$	$\alpha=.025$	$\alpha=.001$
10	0.119571	0.126977	0.130121	0.133258
20	0.0595	0.067185	0.071283	0.079431
30	0.034311	0.039827	0.044452	0.048512
50	0.0186	0.021094	0.023273	0.027471
100	0.009136	0.010902	0.012289	0.01373
200	0.004382	0.005192	0.005598	0.006484
400	0.002197	0.002526	0.002819	0.003242

RAND R. WILCOX

Table 5: Some properties of the g-and-h distribution

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

Table 6: Estimated Probability, $\hat{\alpha}$, of a Type I error, n=20, p=4, $\gamma=.75$, $\rho=.5$

x		$\mathcal{E}$		$\hat{\alpha}$		
g	h	g	h	VP1	VP2	VP3
0.0	0.0	0.0	0.0	0.058	0.057	0.059
0.0	0.0	0.0	0.2	0.052	0.061	0.053
0.0	0.0	0.2	0.0	0.066	0.065	0.048
0.0	0.0	0.2	0.2	0.059	0.063	0.049
0.0	0.2	0.0	0.0	0.041	0.046	0.039
0.0	0.2	0.0	0.2	0.043	0.050	0.045
0.0	0.2	0.2	0.0	0.052	0.065	0.046
0.0	0.2	0.2	0.2	0.046	0.066	0.055
0.2	0.0	0.0	0.0	0.061	0.059	0.058
0.2	0.0	0.0	0.2	0.049	0.051	0.070
0.2	0.0	0.2	0.0	0.047	0.071	0.062
0.2	0.0	0.2	0.2	0.054	0.064	0.054
0.2	0.2	0.0	0.0	0.050	0.056	0.053
0.2	0.2	0.0	0.2	0.042	0.045	0.047
0.2	0.2	0.2	0.0	0.045	0.059	0.049
0.2	0.2	0.2	0.2	0.045	0.060	0.049

TEST OF INDEPENDENCE VIA QUANTILES SENSITIVE TO CURVATURE

Conclusion

One of the main points is that when dealing with the quantiles, the method considered here continues to perform well in simulations, in terms of Type I errors, when sampling from skewed distributions, in contrast to the wild bootstrap method in Stute et al. (1998). Given the ease and flexibility of the method, all indications are that it has practical value. For situations where interpolation is not possible based on the results in Tables 1-4, simulations are still needed to determine critical values, but the results reported here indicate that this needs to be done only once. That is, given n , p and γ , critical values can be determined via simulations, stored in a table, and then used in future studies where these values for n , p and γ occur. An R and S-plus function for applying the method (called medind) is available from the author upon request.

Finally, the modification considered here can be extended to the situation where the goal is to test the fit of a linear quantile regression model. Preliminary results indicate that alternative critical values are now needed and that now critical values have an approximately linear association with $n^{-1.5}$ rather than $1/n$, as was the case here.

References

- Boik, R., & Haaland, B. (2006). Second-order accurate inference on simple, partial, and multiple correlations. *Journal of Modern Applied Statistical Methods*, 5, 283-308.
- Coakley, C. W. & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*, 88, 872-880.
- Dielman, T., Lowry, C. & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics-Simulation and Computation*, 23, 355-371.
- Duncan, G. T., & Layard, M. W. (1973). A monte-carlo study of asymptotically robust tests for correlation. *Biometrika*, 60, 551-558.
- Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician*, 43, 50-54.
- Geiser, P. W. & Randles, R. H. (1997). A nonparametric test of independence between two vectors. *Journal of the American Statistical Association*, 92, 561-567.
- Hafé, A.M.A, Roberts E.M., & Seireg A. A. (1985) Force and velocity during front foot contact in the sprint start. In Winter, D. A. (Ed.) *Biomechanics IX-B*. Champaign, IL: Human Kinetics Publishers, 350-361.
- Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized additive models*. New York: Chapman and Hall.
- Henry, F. M. (1952). Force-time characteristics of the sprint start. *Research Quarterly*, 23, 301-318.
- 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*. New York: Wiley.
- Mammen, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Annals of Statistics*, 21, 255-285.
- Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. New York: Wiley.
- Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, 46, 247-257.
- Reddon, J. R., Jackson, D. N. & Schopflocher, D. (1985). Distribution of the determinant of the sample correlation matrix: Monte Carlo type I error rates. *Journal of Educational Statistics*, 10, 384-388.
- Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379-1389.
- Sievers, W. (1996). Standard and bootstrap confidence intervals for the correlation coefficient. *British Journal of Mathematical and Statistical Psychology*, 49, 381-396.
- Stute, W. Gonzalez Manteiga, W. G., & Presedo Quindimil, M. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association*, 93, 141-149.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, 12, 85-91.
- Wilcox, R. R. (1997). Test of independence and zero correlation among p random variables. *Biometrical Journal*, 39, 183-193.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. (2nd Ed.). San Diego, CA: Academic Press.
- Wilcox, R. R. (2007). On flexible tests of independence and homoscedasticity. *Journal of Modern and Applied Statistical Methods*, 6 30-35.

Regular Articles
**On Measuring the Relative Importance of
Explanatory Variables in a Logistic Regression**

D. Roland Thomas
Carleton University

PengCheng Zhu
Carleton University

Bruno D. Zumbo
University of British Columbia

Shantanu Dutta
St. Francis Xavier University

A search is described for valid methods of assessing the importance of explanatory variables in logistic regression, motivated by earlier work on the relationship between corporate governance variables and the issuance of restricted voting shares (RSF). The methods explored are adaptations of Pratt's (1987) approach for measuring variable importance in simple linear regression, which is based on a special partition of R^2 . Pseudo- R^2 measures for logistic regression are briefly reviewed, and two measures are selected which can be partitioned in a manner analogous to that used by Pratt. One of these is ultimately selected for the variable importance analysis of the RSF data based on its small sample stability. Confidence intervals for variable importance are obtained using the bootstrap method, and used to draw conclusions regarding the relative importance of the corporate governance variables.

Key words: Variable Importance, pseudo-R square, corporate governance.

Introduction

This article describes a search for statistical measures to answer the following applied question: How can one determine the relative importance of correlated explanatory variables in a logistic regression? The case that has motivated this study features a sample of firms listed on the Toronto Stock Exchange, some of which issue restricted voting shares, while the remainder do not (Jog, Zhu, & Dutta 2006).

Firms that have issued restricted shares to the market will henceforth be referred as restricted share firms (RSF) and the combined dataset featuring both types of firms will be referred to as the RSF dataset.

In the case study, logistic regression is used to quantify the relationship between the issuance of restricted voting shares (issue / do not issue) and three constructed measures of corporate governance, namely dispersion of ownership (DISP), suppression of shareholders interests (SUPP) and board independence (INDEP). The methods that will be constructed to assess the relative importance of these explanatory variables will be quite general and can be applied to a wide range of logistic regression problems. The performance of these methods will be evaluated on a constructed dataset that has known properties, and then applied to the RSF dataset. Practitioners frequently ask how to assess variable importance (Healy, 1990), and when the question relates to explanatory variables in logistic regression, the usual recommendation is to inspect the relative magnitudes of the Wald statistics for individual

D. Roland Thomas is Professor Emeritus at the Sprott School of Business, Carleton University, Ottawa, Canada. Email him at rthomas@sprott.carleton.ca. PengCheng Zhu is an Assistant Professor of Finance at the Eberhardt School of Business, University of the Pacific Stockton campus. Shantanu Dutta is an Assistant Professor of Finance at University of Ontario Institute of Technology. Bruno D. Zumbo is a Professor at the University of British Columbia. E-mail: bruno.zumbo@ubc.ca.

explanatory variables (or their square roots which can be interpreted as large sample z -statistics). The problem with this and related approaches can be easily explained with reference to the governance example. For the explanatory variable DISP, its Wald statistic (or its square root z -statistic) shown in Table 3 is a measure of the contribution of DISP to the logistic regression, over and above the contribution of explanatory variables SUPP and INDEP.

Similarly, the Wald statistic for variable SUPP measures its contribution over and above variables DISP and INDEP. Clearly, it is not appropriate to use these two Wald statistics as measures of the relative contribution of DISP and SUPP because the reference set of variables is different in both cases (SUPP and INDEP in the first case, and DISP and INDEP in the second case). The equivalent problem occurs in linear regression, i.e., the t -statistics (or corresponding p -values) for individual variables are not appropriate for assessing relative importance. Considerable attention has been paid to the problem of variable importance in linear regression, evidenced by the work of Pratt (1987), Kruskal (1987), Budescu (1993), Thomas, Hughes and Zumbo (1998), Azen, Budescu and Reiser (2001), Azen and Budescu (2003), Thomas, Zhu, and Decady (2007), and many others.

Although the interpretational questions that arise in logistic regression are generally similar to those encountered in multiple regression (Hosmer & Lemeshow 2000), no comparable attention has been focused on the question of variable importance in the logistic case. The reason for this lack of attention is more likely due to the greater complexity of the logistic model than to any fundamental difference in interpretational requirements. This complexity is also reflected in measures of fit. For example, while R^2 in multiple regression is a widely accepted and natural measure of model fit, which is easily computed and well understood, analogous measures for logistic regression are not as well known. Though several plausible pseudo- R^2 measures have been proposed and compared for logistic regression (Windmeijer 1995; Mittlbock & Schemper

1996), no one measure has yet been accepted as the standard.

The issue of a pseudo- R^2 for logistic regression is particularly relevant to the subject of this paper. One measure of variable importance in multiple regression that has been extensively discussed in the literature is defined in terms of the portion of “total variance explained” that is assigned to each variable. The rule for partitioning R^2 into its individual components, each representing variable importance, was axiomatically justified by Pratt (1987) and has also been given an easily generalized geometric interpretation by Thomas et al. (1998). Thus, to derive a measure of variable importance for logistic regression, it is natural to seek a pseudo- R^2 measure for logistic regression that can be partitioned in an analogous way. It turns out that not all of the pseudo- R^2 measures proposed to date are suitable for such partitioning. A brief review of the better known measures will be given, one of which (Laitila 1993; McKelvey & Zavoina 1975) can be partitioned in a manner similar to that used by Pratt (1987). An additional pseudo- R^2 measure based on a weighted least squares (WLS) representation of the maximum likelihood estimates (MLE) of the logistic regression parameters is also proposed in this paper. This WLS representation lends itself to partitioning using the geometric approach of Thomas et al. (1998), and so provides an alternative set of importance measures, henceforth referred to in this paper as VI indices.

The article is organized as follows. First, the RSF example and dataset are described, along with results of the basic logistic regression analysis. Also described is a large synthetic dataset with population characteristics designed to mimic the sample data, and which will be used throughout to illustrate the properties of the various methods, and to guide the interpretation of the corporate governance case. Next, Pratt’s (1987) axiomatically derived measure of importance for multiple regression is discussed, which will provide the basis for the various sets of VI indices developed in this paper. Specific attention will be paid to the geometric interpretation given by Thomas et al. (1998). Then, a brief account is given of the pseudo- R^2 measures proposed to date for logistic

regression, which (except for the method proposed by Laitila, 1993, and McKelvey & Zavoina 1975) cannot be partitioned using either the axiomatic or the geometric approach.

The pseudo- R^2 measure based on the WLS representation of the logistic MLE is then described. VI indices for logistic regression based on the two pseudo- R^2 measures that can be partitioned are then derived, and their particular characteristics are illustrated using the synthetic dataset. Next, these VI indices are used to shed light on the relative importance of the three governance variables, DISP, SUPP and INDEP. This section also describes the bootstrapping

techniques used to determine standard errors and confidence intervals for VI indices, which are then used to determine the final variable importance orderings. Finally, an overview and recommendations for future research are given.

Example Datasets

Restricted Shares and Corporate Governance

Restricted shares are a regular feature of the Canadian stock market, and unlike traditional common shares which usually carry one voting right per share, restricted shares have reduced voting rights and in some cases carry no voting rights at all. The issuance of restricted

Table 1. Definition of Study Variables

Variables	Explanation
EXPAY	CEO excess payment
BOARD_SIZE	Size of company board of Directors
P_INS_DIR	Percentage of internal Directors on company board
CEO_CHAIR	If CEO is the Chairman of the board (Yes is 1, No is 0)
DIR_OWN	Percentage of Director ownership
DIR_VOT	Percentage of Director voting rights
COM_OWN	Percentage of combined Director and Block ownership
COM_VOT	Percentage of combined Director and Block voting rights
DIR_OWN_VOT	Ratio of Director voting rights to Director ownership

Table 2. Results of the Factor Analysis

Component Score Coefficient Matrix				
	Component			Component Name
	1	2	3	
COM_OWN	0.331	-0.132	-0.159	Dispersion of Ownership and Voting Rights
DIR_OWN	0.311	-0.094	-0.031	
COM_VOT	0.252	0.131	0.061	
DIR_VOT	0.247	0.130	0.143	
DIR_OWN_VOT	-0.071	0.357	0.272	Suppression of Shareholders' Interests
EXPAY	0.006	0.337	-0.145	
BOARD_SIZE	0.017	0.335	-0.150	
P_INS_DIR	0.002	0.000	0.510	Board Independence
CEO_CHAIR	-0.034	0.053	0.517	

Note: Extraction Method: Principal Component Analysis. Rotation Method: Oblimin with Kaiser Normalization

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

shares to the public market reduces the access of non-management shareholders to shares that carry normal voting rights, so that a small number of shareholders (primarily the management group) can effectively control the corporate board. Increasing interest and concern about corporate governance mechanisms in RSFs is now being expressed not only by academic researchers but also by professionals and legislators, particularly in view of the many recent corporate scandals in North America. One of the many objectives of Jog, Zhu and Dutta's (2006) study was to examine the relationship between various corporate governance characteristics and a firm's propensity to issue restricted shares. The final dataset for analysis contained 95 Canadian firms that had restricted shares outstanding on the Toronto Stock Exchange (TSX) between September 1993 and December 2004. A comparison sample was randomly selected from among the TSX companies that had issued no restricted shares during those ten years, providing a combined RSF dataset of 202 firms. A variety of corporate finance and governance variables were collected, as catalogued in Table 1, and a preliminary analysis (not shown) showed the corporate governance variables to be significantly correlated.

A factor analysis and a non-orthogonal "oblimin" rotation was carried out to provide a more succinct and interpretable representation of the variables of Table 1. From Table 2 it can be seen that a useful data summary is provided by three rotated corporate governance factors mentioned in the introduction, namely dispersion of ownership (DISP), suppression of shareholders interests (SUPP) and board independence (INDEP). The estimated correlations between these composites are: (DISP, SUPP) = .06; (DISP, INDEP) = .21 and (SUPP, INDEP) = -.07. Using the SPSS program, scores for each of the corporate governance composite variables were generated using the "regression" method, and saved for subsequent logistic analysis. It should be noted that, in this analysis, no allowance is made for measurement errors arising from the estimation of governance variables that could be regarded as latent. The sampling plan for the Jog et al. (2006) dataset comprises a case-control sample,

in which all RSF firms but only a fraction of the non-RSF firms were sampled. However, it is well known (see Hosmer & Lemeshow 2000, p. 178-181) that when the RSF indicator is treated as a binary random variable, consistent regression parameter estimates are obtained for the explanatory variables; only the estimate of the intercept parameter being inconsistent (or biased).

Because Pratt's (1987) variable importance measures do not depend on the intercept parameter, the case-control nature of the sample will not be a problem. Basic results for the logistic regression of the RSF indicator (RSF=1, non-RSF = 0) on the three composite governance variables are shown in Table 3. A Hosmer-Lemeshow goodness-of-fit test suggests that the model does fit the data ($p = 0.31$).

A Synthetic Dataset

A large synthetic dataset containing 50,000 observations was randomly drawn from a population model designed to partially mimic the corporate governance example. The model features three explanatory variables, with regression parameters equal to the MLEs shown in Table 3, and with explanatory variable means and model covariance matrix set equal to the sample means and sample covariance matrix of the three corporate governance variables. Details of the probabilistic structure of the model, which generates samples that are exactly consistent with a logistic regression model, will be given later. The synthetic dataset will be used to compare the various pseudo- R^2 and corresponding sets of VI indices that will be developed, free of the idiosyncrasies typically present in real data. This will facilitate the interpretation of the new measures when they are applied to the RSF data.

Pratt's Measure of Variable Importance for Multiple Linear Regression

The methods used for developing the variable importance measures for logistic regression will all be adaptations of Pratt's (1987) linear regression method which comprises a particular partition of R^2 . Pratt's method will be outlined in this section given its central importance to the study. A more detailed

summary of Pratt's method is given by Thomas, Zhu, and Decady (2007).

The Axiomatic Approach

Pratt (1987) considered a linear regression equation of the form

$$y = b_0 + b_1x_1 + \dots + b_px_p + u \quad (1)$$

where the disturbance term u is uncorrelated with $x_1, \dots, x_p$, and is distributed with mean zero and variance σ^2 . The total (standardized) population variance, R_p^2 , explained by model (1) can be written as

$$R_p^2 = \sum_j \beta_j \rho_j \quad (2)$$

where β_j is the usual standardized regression coefficient corresponding to x_j , and ρ_j is the simple correlation between y and x_j . Pratt justified the rule whereby relative importance is equated with variance explained, provided that explained variance attributed to x_j is $\beta_j \rho_j$. This definition of variable importance has been widely used in the applied literature (Green,

Carroll and De Sarbo 1978), but as documented by Pratt (1987), it has also been severely criticized. Pratt justified the measure using an axiomatic approach based largely on symmetry and invariance to linear transformation. Subject to his axioms, he showed that his measure is unique. An added bonus is that Pratt's measure allows the importance of a subset of variables to be defined additively, as the sum of their individual importances. Other commonly used measures do not allow for an additive definition.

The Geometric Approach

Thomas et al. (1998) gave a sample interpretation of Pratt's measure based on the geometry of least squares. They considered a sample of N observations fitted to a model of the form (1), so that the observed variables $y, \mathbf{x}_1, \dots, \mathbf{x}_p$ comprise vectors in an N -dimensional space. Without loss of generality

they assumed that all variables have zero mean, i.e., $\mathbf{y}'\mathbf{I}_N = \mathbf{x}_1'\mathbf{I}_N = \dots = \mathbf{x}_p'\mathbf{I}_N = 0$, where $\mathbf{I}_N$ is

an $N \times 1$ vector of ones. In this case, $\hat{\mathbf{y}}$, the fitted value of $\mathbf{y}$, is the projection of $\mathbf{y}$ onto the subspace spanned by the explanatory variables

Table 3

Logistic Regression Results for the Combined RSF Dataset

	$\tilde{\mathbf{b}}$	s.e. ($\tilde{\mathbf{b}}$)	Wald	df	$\exp(\tilde{\mathbf{b}})$
Intercept	0.196	0.236	0.69	1	1.127
DISP	1.290	0.252	26.30	1	3.633
SUPP	2.495	0.397	39.53	1	12.120
INDEP	0.915	0.238	14.78	1	2.497

Note: Included in this table is the value of $\exp(\tilde{\mathbf{b}}_j)$, $j = 1, 2, 3$, where $\tilde{\mathbf{b}}_j$ denotes the MLEs of the logistic regression coefficient for the j 'th of the three explanatory variables. The exponential of the j 'th regression parameter represents the proportional increase in the odds of a firm being an issuer of restricted voting shares corresponding to an increase of one unit in its score on the j 'th explanatory variable, with all other scores held constant. While it is tempting to use these odds ratios as measures of relative importance, it is easily seen that they suffer from precisely the same flaw as do the Wald or z-statistics.

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

$\mathbf{x}_j, j = 1, \dots, p$, and has the representation

$$\hat{\mathbf{y}} = \hat{b}_1 \mathbf{x}_1 + \dots + \hat{b}_p \mathbf{x}_p, \quad (3)$$

where the $\hat{b}_j$'s are least squares estimates of the population regression coefficients

$b_j, j = 1, \dots, p$. Figure 1 illustrates the geometric interpretation of Pratt's importance measures in a two-variable model subspace. In this model subspace, appropriate multiples of $\mathbf{x}_1$ and $\mathbf{x}_2$ (given by the least squares

estimates of the regression coefficients, $\hat{b}_1$ and $\hat{b}_2$, respectively) sum geometrically to $\hat{\mathbf{y}}$, the projection of $\mathbf{y}$ from its N -dimensional space onto the two-dimensional model subspace. The heavy lines represent the vector projection of each component $\hat{b}_j \mathbf{x}_j$ onto $\hat{\mathbf{y}}$. Clearly, the orthogonal components sum to zero. Thus it is

natural to use the (signed) lengths of the individual projections in the $\hat{\mathbf{y}}$ direction (which sum to $\hat{\mathbf{y}}$) as measures of the contribution of each $\mathbf{x}_j$ to $\hat{\mathbf{y}}$, i.e., as measures of variable importance. Thomas et al. (1998) actually defined their VI indices, denoted d_j , as the ratio of the signed length of these projections to the length of $\hat{\mathbf{y}}$, and showed that

$$d_j = \hat{\beta}_j \hat{\rho}_j / R^2, j = 1, \dots, p, \quad (4)$$

where hats denote sample estimates, and where R^2 is the usual proportion of sample variance explained. It can be seen that the VI indices defined in equation (4) are sample estimates of Pratt's (1987) measures, normalized by R^2 . Defined in this way, they automatically sum to one. The d_j 's are analogous to the

Figure 1 Importance Measures as Projections

discriminant ratio coefficients (DRC's) introduced by Thomas (1992) as variable importance measures for descriptive discriminant analysis.

Negative Values of Pratt's VI Indices

Pratt's measure can be negative, a feature that has been criticized and that would appear to detract from its utility as a measure of importance. However, according to Pratt's axiomatic derivation, the importance rule is valid only when the population quantities $\beta_j \rho_j$ are all positive. Thus negativity of any one of these quantities does not signify negative importance, but instead signifies a regression situation which is "too complex for a single measure" (Pratt 1987, p. 245). Thomas et al. (1998) used an extension of the geometric argument to show that negative d_j 's of large magnitude can arise from multicollinearity among the predictor variables. They gave an example where a negative VI index of large magnitude (close to one) was reduced to a small positive value by the application of ridge regression (Hoerl & Kennard 1970), suggesting that the original "negative importance" was false. Not all negative importances will be false, however, and the fact must be faced that some regression modeling situations are so complex that there is no single measure of variable importance that satisfies Pratt's axioms. For multiple linear regression, Thomas, Zhu and Decady (2007) have developed simultaneous confidence interval procedures that can be used to identify such cases.

Pratt's axiomatic derivation provides a theoretical foundation for his measure in the case of multiple regression, but it is not necessarily easy to generalize his method to other analyses. The benefit of the geometric interpretation is that it is sometimes easier to apply to other modeling techniques than is the axiomatic approach, as exemplified by Thomas's (1992) introduction of DRC's in discriminant analysis. It will be shown in Section 5 that both the axiomatic and geometrical interpretations of Pratt's method can be extended to the case of logistic regression.

R^2 Measures for Logistic Regression

The Model Setup

The logistic model of interest can be expressed as

$$\log [\pi_i / (1 - \pi_i)] = \mathbf{x}_i' \mathbf{b},$$

$$i = 1, \dots, N, \quad (5)$$

where $\pi_i = P(y_i = 1 | \mathbf{x}_i)$, and where in this logistic case, $y_i, i = 1, \dots, N$ are independent binary random variables, $\mathbf{x}_i, i = 1, \dots, N$ are $(p+1)$ -vectors of observed explanatory variable scores (with first element equal to one) for the i 'th individual, and $\mathbf{b}$ is a $(p+1)$ -vector of regression coefficients (with first element b_0 corresponding to the intercept).

The reader is warned not to confuse the $\mathbf{x}_i, i = 1, \dots, N$, notation used in equation (5), which refers to N sample realizations of a $(p+1)$ -vector, with the notation $\mathbf{x}_j, j = 1, \dots, p$, used in the previous section, which referred to p realizations of an N -vector. The indexing will always be clearly specified to avoid confusion. Also, no notational distinction is made in the paper between a random variable and its realization; the distinction will be clear from the context. In equation (5) it will be assumed that at least one of the predictors will be measured on a continuous scale, so that none of the covariate patterns will be repeated. This is the sparse case in which the Pearson chi-square and the deviance (discussed, for example, by McCullagh and Nelder 1989) do not exhibit their "usual" asymptotic chi-squared distributions, and for which appropriate goodness-of-fit measures are still an issue. The aim of this section is to identify, for the above setup, measures of fit of the R^2 type that can be partitioned to yield VI indices for logistic regression. Some of the relevant R^2 measures proposed to date will be briefly reviewed.

Pseudo- R^2 Measures for Logistic Regression

In a review of pseudo- R^2 measures for binary choice models, Windmeijer (1995) reviewed several categories of measures of fit, the first of which is usually attributed to Efron (1978) though it has been considered by a number of authors. It has the form

$$R_E^2 = 1 - \frac{\sum_{i=1}^N (y_i - \tilde{\pi}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (6)$$

where $\bar{y}$ is the sample mean of the binary y_i 's and the $\tilde{\pi}_i$'s in this case denote maximum likelihood estimates (MLEs) of the π_i 's. In fact, any consistent estimates of the π_i 's will suffice. Mittlbock and Schemper (1996) favored this measure over many of its competitors. However, Cameron and Windmeijer (1996) noted that the lower bound for Efron's measure is not in general equal to zero, and may in some cases be negative. For this reason, and because it cannot readily be partitioned to identify the contribution of individual predictor variables, Efron's measure will not be considered further.

The second category consists of measures based on the loglikelihood corresponding to model (5), namely

$$\begin{aligned} & \log L(\mathbf{b}) \\ &= \sum_i [y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)]. \end{aligned} \quad (7)$$

McFadden's (1974) measure has the form

$$R_{MF}^2 = 1 - \log L(\tilde{\mathbf{b}}) / \log L_0, \quad (8)$$

where $L(\tilde{\mathbf{b}})$ denotes the likelihood evaluated at the maximum likelihood estimate $\tilde{\mathbf{b}}$, and L_0 denotes the likelihood for the model containing only an intercept term. When there are no repeated predictor patterns, R_{MF}^2 lies in the interval $[0, 1]$. Otherwise, its upper limit is less than one, in which case the statistic can be adjusted to recover the appropriate limits (Hosmer & Lemeshow 2000, pp. 164). McFadden's measure possesses several attractive features. It is related to the asymptotic chi-squared test that a subset of the model parameters are zero, and it also has an information theoretic interpretation (see

Windmeijer, 1995). Unfortunately, it cannot be partitioned into individual importances, either by means of the linear geometric interpretation described earlier or by any other means known to the authors. A related measure due to Cox and Snell (1989) is also based on the likelihood ratio, and has the form

$$R_{CS}^2 = 1 - [L_0 / L(\tilde{\mathbf{b}})]^{2/N}. \quad (9)$$

This measure does not attain an upper limit of one when the model fits perfectly, and it was suggested by Cragg and Uhler (1970) that it should be scaled to give the required upper bound. Nagelkerke (1991) advocated the same scaling and showed that the scaled measure possesses theoretically attractive features. However, Mittlbock, and Schemper (1996) criticized this scaling as cosmetic, noting that there is no theoretical reason why such a scaling should be appropriate at intermediate values of the statistic. As with the McFadden measure, there appears to be no way to partition R_{CS}^2 or its scaled counterpart to account for contributions of individual variables.

A third category of R^2 measures is based on the interpretation of logistic regression (and other binary choice models) as a linear regression of predictors on an unobservable continuously distributed random variable y_i^* , where the observed binary variable y_i takes the value 1 for $y_i^* \geq 1$, and the value 0 for $y_i^* < 1$. The linear model is specified as

$$y_i^* = \mathbf{x}_i' \mathbf{b} + \varepsilon_i, \quad i = 1, \dots, N \quad (10)$$

where the ε_i are independently distributed logistic variables with mean zero and variance $\pi^2/3$, with $\mathbf{x}_i'$ and $\mathbf{b}$ defined as in equation (10).

Had the response variable $\mathbf{y}^* = (y_1^*, \dots, y_N^*)'$ been observed, then standard OLS parameter and residual estimation could be used resulting in a measure R^{*2} . Although $\mathbf{y}^*$ is not observable, R^{*2} can be replaced by a measure proposed by McKelvey and Zavoina (1975) and

Laitila (1993), henceforth referred to as MZL, namely

$$R_{MZL}^2 = \hat{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{X} \hat{\mathbf{b}} / (\hat{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{X} \hat{\mathbf{b}} + N\pi^2 / 3). \quad (11)$$

Here $\hat{\mathbf{b}}$ is any consistent estimator of $\mathbf{b}$, $\mathbf{X}$ is an $N \times (p+1)$ matrix having rows $\mathbf{x}_i'$, $i=1, \dots, N$, and $\mathbf{Q}$ is the $N \times N$ projection matrix given by $\mathbf{Q} = \mathbf{I}_N - \mathbf{I}_N \mathbf{I}_N' / N$. Laitila (1993) gave a more general version of (11) applicable also to limited dependent variable models, in which the error variance term was consistently estimated. The key property of R_{MZL}^2 (and its more general versions) is that the difference between it and R^{*2} vanishes with increasing sample size, i.e., it is asymptotically zero in probability. It will be shown in the next section that Pratt's approach can be applied to partition R_{MZL}^2 to yield a set of VI indices. It will also be shown that even though the original R^{*2} itself is unobservable, it nevertheless provides the basis for deriving an alternative set of normalized VI indices that are asymptotically close to those derived from R_{MZL}^2 .

An R^2 Measure Based on Weighted Least Squares

It was noted by Pregibon (1981) that the maximum likelihood estimator $\tilde{\mathbf{b}}$ of $\mathbf{b}$ can be represented in terms of the weighted least squares regression of a vector of pseudo-values $\mathbf{z}$ on $\mathbf{X}$, given by

$$\tilde{\mathbf{b}} = (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V} \mathbf{z}, \quad (12)$$

where $\mathbf{z} = \mathbf{X} \tilde{\mathbf{b}} + \mathbf{V}^{-1} \mathbf{r}$, $\mathbf{r} = (\mathbf{y} - \tilde{\boldsymbol{\pi}})$, $\mathbf{y}$ is the

$N \times 1$ vector of binary observations, $\tilde{\boldsymbol{\pi}}$ is the $N \times 1$ vector of estimated probabilities corresponding to the maximum likelihood estimate $\tilde{\mathbf{b}}$, and $\mathbf{V}$ is the $N \times N$ diagonal weight matrix having elements $\tilde{\pi}_i (1 - \tilde{\pi}_i)$, $i = 1, \dots, N$. Pregibon (1981) exploited equation

(12) to extend the diagnostic techniques of linear regression to logistic regression, and Nordberg (1981) and Hosmer, Jovanovic, and Lemeshow (1989) used it to apply the techniques of all subsets variable selection to logistic regression. In this section the representation (12) will be used to develop a pseudo- R^2 measure for logistic regression.

It will be more convenient to represent equation (12) as the OLS regression of $\boldsymbol{\omega} = \mathbf{V}^{1/2} \mathbf{z}$ on $\mathbf{V}^{1/2} \mathbf{X}$, with fitted values $\hat{\boldsymbol{\omega}} = \mathbf{V}^{1/2} \mathbf{X} \tilde{\mathbf{b}}$ and residuals given by

$$\boldsymbol{\omega} - \hat{\boldsymbol{\omega}} = \mathbf{V}^{1/2} (\mathbf{z} - \mathbf{X} \tilde{\mathbf{b}}) = \mathbf{V}^{-1/2} \mathbf{r} \quad (13)$$

The residual sum of squares from this pseudo-regression is

$$\begin{aligned} SS_E &= (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}})' (\boldsymbol{\omega} - \hat{\boldsymbol{\omega}}) = \mathbf{r}' \mathbf{V}^{-1} \mathbf{r} \\ &= \sum_{i=1}^N \frac{(y_i - \tilde{\pi}_i)^2}{\tilde{\pi}_i (1 - \tilde{\pi}_i)} = \chi^2, \end{aligned} \quad (14)$$

the familiar Pearson "chi-squared" statistic. Alternatively,

$$SS_E = \boldsymbol{\omega}' \mathbf{M} \boldsymbol{\omega} = \mathbf{z}' \mathbf{V}^{1/2} \mathbf{M} \mathbf{V}^{1/2} \mathbf{z} \quad (15)$$

where $\mathbf{M}$ is a $N \times N$ projection matrix of rank $N-p-1$ given by

$$\mathbf{M} = \mathbf{I} - \mathbf{V}^{1/2} \mathbf{X} (\mathbf{X}' \mathbf{V} \mathbf{X})^{-1} \mathbf{X}' \mathbf{V}^{1/2}. \quad (16)$$

This projection matrix, derived from the weighted least squares representation of the maximum likelihood estimate, was used by Pregibon (1981) in his development of logistic regression diagnostics. The OLS version of the maximum likelihood identity also yields a regression sum of squares, given by

$$\begin{aligned} SS_R &= \hat{\boldsymbol{\omega}}' \hat{\boldsymbol{\omega}} - \hat{\boldsymbol{\omega}}' \hat{\boldsymbol{\omega}} \mathbf{V}^{1/2} \mathbf{I} (\mathbf{I}' \mathbf{V} \mathbf{I})^{-1} \mathbf{I}' \mathbf{V}^{1/2} \hat{\boldsymbol{\omega}} \\ &= \tilde{\mathbf{b}}' \mathbf{X}' \mathbf{V}^{1/2} \mathbf{Q}_V \mathbf{V}^{1/2} \mathbf{X} \tilde{\mathbf{b}}, \end{aligned} \quad (17)$$

where $\mathbf{Q}_V = \mathbf{I} - \mathbf{V}^{1/2} \mathbf{I} (\mathbf{I}' \mathbf{V} \mathbf{I})^{-1} \mathbf{I}' \mathbf{V}^{1/2}$. Note that equation (17) comprises a weighted version of the numerator of R_{MZL}^2 given in equation

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

(11). Equations (14) and (17) immediately lead to an R^2 measure given by

$$R_{WLS}^2 = SS_R / (SS_R + \chi^2) \quad (18)$$

$$= 1 - \chi^2 / (SS_R + \chi^2). \quad (19)$$

The geometric interpretation of Pratt's measures will be used in the next section to partition R_{WLS}^2 and yield a set of normalized VI indices.

A Numerical Comparison of the Pseudo- R^2 Measures

Pseudo- R^2 measures for the synthetic and RSF datasets are displayed in Table 4. Results for the synthetic data, shown in the leftmost column, can be regarded as population values essentially free of sampling error. Values of Efron's R_E^2 , McFadden's R_{MF}^2 and Cox and Snell's R_{CS}^2 are shown for reference only, as they cannot be partitioned and thus do not provide the basis for the development of VI indices. As is typical of such pseudo- R^2 measures, they vary considerably in magnitude (Mittlbrock & Schemper 1996) for both datasets.

Of all the measures in Table 4, the largest value is recorded by R_{MZL}^2 (McKelvey & Zavoina 1975; Laitila 1993), which is not surprising because it is designed to measure the explained variation in the continuous latent variable $\mathbf{y}^*$, rather than the variation in the observed vector of binary variables $\mathbf{y}$. On the other hand, the new weighted least squares measure R_{WLS}^2 records the smallest pseudo- R^2 value of all, for both synthetic and RSF data. It is interesting to note that Mittlbrock and Schemper (1996) argued against using the weighted least squares representation of the MLE to construct a pseudo- R^2 because of the potentially distorting effect of the weights. Generally speaking, the trends exhibited in Table 4 for pseudo- R^2 values are similar for both the RSF dataset and the synthetic dataset, which represents population values. The sample values obtained for the RSF dataset can therefore be validly used for interpretational

purposes, even though the sample size is not large.

Variable Importance Indices for Logistic Regression Measures of Importance Based on R_{MZL}^2 and R^{*2}

The continuous model (10) satisfies the assumptions of Pratt's axiomatic approach to variable importance for linear models, the only difference being that the dependent variable y^* is not observable. Thus the VI indices of equation (4) can be applied provided only that consistent estimates of β_j , ρ_j and R^2 can be obtained. An estimate of R^2 is given by R_{MZL}^2 , as described in the previous section, and a consistent estimate of β_j is given by

$$\hat{\beta}_j^{MZL} = \tilde{b}_j \hat{\sigma}_j / \sqrt{\tilde{b}' \mathbf{X}' \mathbf{Q} \mathbf{X} \tilde{b} + N \pi^2 / 3}, \quad (20)$$

(see equation 11) where $\tilde{b}_j$ is the (known) MLE of the regression coefficient for the j 'th predictor variable x_{ij} , $i=1, \dots, N$, and where $\hat{\sigma}_j^2 = [\mathbf{X}' \mathbf{Q} \mathbf{X}]_{jj}$ is its sample variance.

The correlation ρ_j between $\mathbf{y}^*$ and each observed predictor $\mathbf{x}_j$, $j=1, \dots, p$ can be estimated as a polyserial correlation, $\hat{\rho}_j^{PS}$ (Dragow 1986), inferred using only the observed binary responses y_i , $i=1, \dots, N$ and the observed predictors. These estimates together yield the set of VI indices

$$d_j^{MZL} = \hat{\beta}_j^{MZL} \hat{\rho}_j^{PS} / R_{MZL}^2, \quad j=1, \dots, p. \quad (21)$$

Note, however, that this application of polyserial coefficients invokes an assumption of joint multivariate normality of y_i^* and $\mathbf{x}_i$ which is not required in the development of R^{*2} or R_{MZL}^2 . Further, since ε_i in equation (10) is assigned a logistic distribution, y_i^* itself will

not be multivariate normal. Calculations based on the synthetic dataset yield $R_{MZZ}^2 = 0.741$, and $\sum_j \hat{\beta}_j^{MZL} \hat{\rho}_j^{PC} = 0.747$, indicating that for normal $\mathbf{x}_i$'s and logistic ε_i , estimates of the polyserial correlations are robust to this violation of joint normality when the predictors themselves are normal. (The corresponding comparison for the RSF data yields 0.737 versus 0.758). Despite this robustness, it is nevertheless worth seeking normalized VI indices that do not rely on polyserial correlation estimates.

An alternative expression for VI indices can be obtained by applying the derivation of Thomas, Hughes and Zumbo (1998) with $\mathbf{y}^*$ treated as known. This yields

$$d_j^* = [\mathbf{y}^{*'} \mathbf{Q} \mathbf{x}_j \hat{b}_j / N] / R^{*2} = [\hat{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{x}_j \hat{b}_j / N] / R^{*2} \quad (22)$$

where $\hat{\mathbf{b}}'$ and $\hat{b}_j$ represent OLS regression parameter estimates. Thus knowledge of $\mathbf{y}^*$ is not needed to define VI indices; consistent estimates of the population values of d_j^* can be obtained by replacing $\hat{\mathbf{b}}$ by the MLE $\tilde{\mathbf{b}}$, and R^{*2} by R_{MZZ}^2 . As a result of these replacements, the sum of the d_j^* 's will sum to one asymptotically, without the slight approximation inherent in the method that relies on the polyserial coefficient. Furthermore, normalized VI indices that sum identically to one can be defined as

$$d_j^*(N) = \frac{\tilde{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{x}_j \tilde{b}_j}{\sum_j \tilde{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{x}_j \tilde{b}_j} = \frac{\tilde{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{x}_j \tilde{b}_j}{\tilde{\mathbf{b}}' \mathbf{X}' \mathbf{Q} \mathbf{X} \tilde{\mathbf{b}}}, \quad (23)$$

where the denominator of equation (23), divided by N, is asymptotically equivalent to R^{*2} and R_{MZZ}^2 . Equation (23) represents the most convenient version of a VI index based on the linear representation (10). Values of both d_j^{MZL}

and $d_j^*(N)$ for the synthetic dataset are displayed in Table 5.

Measures of Importance Based on R_{WLS}^2

The assumptions underlying Pratt's axiomatic approach do not apply to the WLS representation of the MLE given in equation (12). However, the measure of fit R_{WLS}^2 can be partitioned by applying the geometric approach of Thomas et al. (1998) to the pseudo-regression formulation of Section 4, i.e., by an appropriate interpretation of equation (4). Let $\tilde{\beta}_j$ represent the standardized logistic regression coefficient corresponding to the j th predictor, $j = 1, \dots, p$, given by

$$\tilde{\beta}_j = b_j (x_j' V^{1/2} Q_V V^{1/2} x_j)^{1/2} / (z' V^{1/2} Q_V V^{1/2} z)^{1/2} \quad (24)$$

where $\tilde{b}_j$ is the maximum likelihood estimate of the j th logistic regression coefficient b_j , and let $\tilde{\rho}_j$ be the correlation between $\omega = V^{1/2} z$ and $V^{1/2} x_j$, given by

$$\tilde{\rho}_j = \frac{z' V^{1/2} Q_V V^{1/2} x_j}{\left\{ (z' V^{1/2} Q_V V^{1/2} z)^{1/2} (x_j' V^{1/2} Q_V V^{1/2} x_j)^{1/2} \right\}} \quad (25)$$

Then the required VI indices for the j 'th predictor variable (i.e. the j 'th partition of R_{WLS}^2) are obtained from equation (4) as

$$d_j^{WLS} = \tilde{\beta}_j \tilde{\rho}_j / R_{WLS}^2. \quad (26)$$

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

Table 4. Pseudo- R^2 Measures for the Synthetic and Restricted Share Firms Datasets

Pseudo R^2 Measures	Synthetic Data (N=50,000)	RSF Data (N = 202)
R_{MZL}^2 (equation 11)	0.741	0.737
R_{WLS}^2 (equation 18)	0.193	0.226
R_E^2 (equation 6)	0.549	0.483
R_{MF}^2 (equation 8)	0.487	0.507
R_{CS}^2 (equation 9)	0.491	0.504

Table 5. Variable Importance Indices for the Synthetic and RSF Datasets

VI Indices		Synthetic Data (N=50,000)			RSF Data (N = 263)		
		DISP	SUPP	INDEP	DISP	SUPP	INDEP
d_j^{MZL}	(equation 21)	.226	.683	.096	.297	.619	.113
$d_j^*(N)$	(equation 23)	.224	.680	.096	.227	.674	.099
d_j^{WLS}	(equation 26)	.224	.682	.094	.374	.499	.127
$d_j^{WLS}(N)$	(equation 27)	.224	.682	.094	.374	.499	.127

In practice, with dependent variable $V^{1/2}z$ and predictor variables $V^{1/2}X$, the quantities $\tilde{\beta}_j, j = 1, \dots, p$, and R_{WLS}^2 can be obtained from the output of standard multiple regression programs as the standardized regression coefficient (the “beta” weight in SPSS, for example) and the standard R^2 measure, respectively. Similarly, the correlation $\tilde{\rho}_j$ corresponds to the standardized regression coefficient in the simple linear regression of $V^{1/2}z$ on $V^{1/2}X$.

An algebraically equivalent representation of d_j^{WLS} can be derived in a manner similar to that used to derive equation (22), by applying regression identities to the

WLS representation of the MLE $\tilde{b}$ defined by equation (12). This leads to the expression

$$d_j^{WLS}(N) = \frac{\tilde{b}'X'V^{1/2}Q_vV^{1/2}x_j\tilde{b}_j}{\tilde{b}'X'V^{1/2}Q_vV^{1/2}X\tilde{b}}, \quad (27)$$

which yields a weighted least squares analogue of equation (23).

A Numerical Comparison of the Competing VI Indices

Values of the variable importance indices described in the previous section are shown in Table 5 for the three corporate governance variables. The third and fourth rows of the table simply illustrate the fact that equations (26) and (27) are algebraically

equivalent, with VI indices that sum to one. These alternative forms will be referred to in what follows as d_j^{WLS} .

It can be seen from the first row of Table 5 that the VI indices d_j^{MZZ} do not sum to one exactly, as was explained in the text. For the synthetic dataset they sum to 1.008 and for the

RSF dataset they sum to 1.028, both representing only minor discrepancies. The indices $d_j^*(N)$ do sum to one by virtue of their construction, and will be used henceforth in preference to d_j^{MZZ} . Thus the important conclusions to be drawn from Table 5 relate to the two index sets $d_j^*(N)$ and d_j^{WLS} .

For the former, it can be seen that individual indices for the three independent variables are very similar for both the synthetic (population) dataset and the RSF dataset. This suggests that the VI indices $d_j^*(N)$ perform well for moderate size samples, a conclusion that should be explored in greater detail in a more extensive simulation study. It can also be seen from Table 5 that both sets of indices, $d_j^*(N)$ and d_j^{WLS} , exhibit very similar results for the large sample synthetic dataset. However, for the moderately sized RSF dataset, the d_j^{WLS} indices differ noticeably from these large sample values, suggesting that the VI indices d_j^{WLS} might be less robust to small and medium sample sizes than the indices $d_j^*(N)$.

It was noted earlier that Mittlbock and Schemper (1996) recommended against using the weighted least squares representation of the logistic regression MLE because of the potentially distorting effect of the weights. While these weights appear to have little impact on either set of VI indices for the large sample synthetic dataset, their effect may be more severe for smaller sample sizes. For this reason, the following analysis of variable importance in the RSF dataset will be based entirely on the $d_j^*(N)$ indices.

An Analysis of Variable Importance for the RSF Dataset

Point Estimates of Importance

The point estimates of the VI indices $d_j^*(N)$ suggest that SUPP (suppression of shareholders' interests) is the most important governance variable for differentiating between restricted share firms and non-restricted shares firms, and that INDEP (board independence) is the least important, with the effect of DISP (dispersion of ownership) being intermediate. However, to decide if these differences between point estimates translate into real (population) differences in variable importance, standard errors and confidence intervals for each individual index must be estimated. Thomas, Zhu, and Decady (2007) provided large sample formulas for the standard errors of normalized Pratt indices for the linear regression case, but it is not practical to extend their analysis to the logistic regression case. However, because the VI indices proposed in this paper are smooth functions of means, variance and covariances, standard errors can be obtained using the bootstrap resampling methodology, as described in the following section.

Standard Errors and Confidence Intervals for the VI Indices

A standard non-parametric bootstrap (Efron and Tibshirani 1993) was used to estimate the standard errors and corresponding confidence intervals for the indices $d_j^*(N)$. The resampling procedure consisted of 1000 independent bootstrap samples of 200 observations (each taken with replacement from the original RSF sample). From the 1000 bootstrap samples, 1000 replications of the logistic parameter estimates and VI indices were then calculated, allowing for the computation of bootstrap standard errors, as well as a visual depiction of the bootstrap distribution. All computations were carried out using the bootstrap facilities of the R language (Canty and Ripley 2006). Histograms of the bootstrap samples for the VI indices $d_j^*(N)$ are shown in Figure 2, and corresponding bootstrap standard errors are shown in Table 6.

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

Figure 2

Bootstrap Histograms of The VI Indices $d_j^*(N)$

Table 6

Bootstrap Standard Errors and BCa Confidence Intervals for VI Indices $d_j^*(N)$

Variables	Point Estimates	Standard Errors	Individual 95% CIs	Simultaneous 95% CIs
DISP	.227	.077	(.095, .400)	(.072, .434)
SUPP	.674	.087	(.490, .831)	(.440, .841)
INDEP	.099	.044	(.032, .214)	(.018, .232)

Large sample confidence intervals are often computed simply as a point estimate plus and minus twice the standard deviation of the statistic in question. However, in cases where the sampling distribution still retains some non-normality, such confidence intervals tend to provide poor coverage. Numerous alternatives based on the bootstrap have been described in the literature (Efron and Tibshirani 1993; Davison and Hinkley 1997), and it has been shown that the Bias Corrected and Accelerated (BCa) interval has superior coverage properties (Platt, Hanley and Yang 2000). A major advantage of the BCa interval is its transformation-respecting property, i.e., the method effectively selects the best (most normal) scale and then transforms the interval back to the original scale of interest (Efron 1987). Individual BCa 95% confidence intervals (CIs) for the VI indices $d_j^*(N)$ are shown in Table 6 along with the point estimates and standard errors.

Individual confidence intervals are appropriate if the VI index of a specific variable is of prior interest. If the interest results from the

analysis itself, i.e., if the VI index of a particular variable is the largest, which implies a comparison with all other variables, then simultaneous confidence intervals should be used (Thomas et al. 2007). As shown by the latter authors, simultaneous confidence intervals can be obtained using the Bonferroni adjustment which, for the RSF case featuring three explanatory variables, implies constructing confidence intervals each at a nominal alpha level of $100(1 - .05/3)\%$. These also are shown in Table 6.

From Table 6 it can be seen that the indices $d_j^*(N)$ yield simultaneous confidence intervals for DISP and SUPP that do not overlap, suggesting that SUPP is more important than DISP, as indicated by the point estimates. Simultaneous confidence intervals for the VI indices for DISP and INDEP do overlap, however, suggesting that the population importances of these two variables may not actually be different. The simultaneous confidence intervals are illustrated graphically in Figure 3.

Conclusions

This article has described a search for variable importance measures appropriate for logistic regression, motivated by earlier work on the relationship between corporate governance variables and the issuance of restricted shares. Two methods have been proposed, both of which are based on Pratt's (1987) axiomatically derived partition of R^2 for multiple linear regression, which can be generalized using the geometric interpretation described by Thomas et al. (1998). The first method uses a pseudo- R^2 measure for logistic regression proposed by McKelvey and Zavoina (1975) and Laitila (1993), which represents a logistic regression as the binary truncation of an unobservable dependent variable that is linearly related to the explanatory variables of interest.

This method yields a set of VI indices denoted $d_j^*(N)$ in the paper. The second method uses a representation of the maximum likelihood estimate of the logistic regression coefficients as a weighted least squares (WLS) regression, a representation exploited earlier by Pregibon (1981), Nordberg (1981) and Hosmer, Jovanovic and Lemshow (1989). A set of VI indices, denoted d_j^{WLS} , are then derived by applying a geometric analogue of Pratt's partitioning approach to the WLS version of R^2 based on this representation. Both sets of indices satisfy the property that they sum to one, which gives each index a meaningful scale, and they also share the property of additivity, namely that the importance of a subset of variables is equal to the sum of their individual importances, a property not shared by competing measures. A large synthetic dataset was constructed to mimic the actual data and was used to explore the small/medium sample properties of the two main methods. The indices $d_j^*(N)$ exhibited more stable small sample behaviour and were therefore used in the final analysis of variable importance.

In the analysis of the motivating case, the VI indices $d_j^*(N)$ were used to assign importances to three corporate governance factors that highlight difference in governance characteristics between firms with restricted

share structure and other public firms without this structure. These variables were SUPP (suppression of shareholders interests), DISP (dispersion of ownership) and INDEP (board independence). A non-parametric bootstrap method was used on the RSF dataset to make statistical inferences on the importance measures.

Standard errors together with individual and simultaneous confidence intervals were estimated for each importance measure of the governance factors in the logistic regression model. The bias corrected and accelerated interval method (BCa) was employed to ensure good coverage performance of the confidence interval (Efron 1987; Platt, Henley and Yang 2000). The inferential analysis revealed that the most important contribution to the logistic regression, i.e., to the probability that a firm will issue restricted voting shares, is made by the variable SUPP. Although point estimates of importance suggest that variable DISP is more important than INDEP, examination of the simultaneous confidence intervals reveals that the importances of these two variables are not significantly different. It can be seen from the earlier results shown in Table 3 that the ranking suggested by the regression coefficients (which have identical scales because of the unit variances of the composite variables) and the Wald statistics are the same for the RSF variables as those suggested by the VI indices. This will not be the same in all situations, however, and occurs in this case because of the relatively small correlations between the explanatory corporate governance variables.

Though the development of the VI indices $d_j^*(N)$ described in this paper was motivated by an analysis of the RSF dataset, these indices and the general methodology can be applied to any logistic regression which can be modeled in terms of an underlying continuous response. Alternatively, if this assumption is deemed untenable in some situation, the alternative VI indices d_j^{WLS} based on the WLS representation can be used. It is important to note, however, that the examination of the properties of both sets of indices has been limited to a comparison with an empirically generated population. Further research involving

simulation studies is needed to examine in detail the small and medium sample biases and confidence interval coverage rates of both sets of indices. In the meantime, however, the theoretical developments described in this paper provide a viable solution to the vexing problem of determining the relative importance of explanatory variables in a logistic regression analysis.

Authors' Note

A preliminary version of this paper was presented at the 2006 Administrative Science Association of Canada (ASAC) conference where it was awarded an honorable mention award in the Management Science track. The authors thank Professor Vijay Jog, Sprott School of Business, Carleton University, for permission to use the data that motivated the proposed methodology. The work of Thomas was supported by a grants from the Natural Sciences and Engineering Research Council of Canada, and from the Canadian research network MITACS (Mathematics of Information Technology and Complex Systems). Work of Zumbo was supported by grants from the Social Sciences and Humanities Research Council of Canada, and the Canadian Institutes for Health Research.

References

- Azen, R., Budescu, D. V., and Reiser, B. (2001). Criticality of predictors in multiple regression. *British Journal of Mathematical and Statistical Psychology*, 54, 201-225.
- Azen, R., and Budescu D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, 8, 129-148.
- Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, 114, 542-551.
- Cameron, A. C. and Windmeijer, F. A. G. (1996). R-squared measures for count data regression models with applications to health-care utilization. *Journal of business and Economics Statistics*, 14, 209-220.
- Canty, A. and Ripley, B. (2006). Bootstrap R (S-Plus) functions. (<http://cran.r-project.org/src/contrib/Descriptions/boot.html>)
- Cox, D. R. and Snell, E. J. (1989). *Analysis of binary data*. 2nd Edition, London: Chapman and Hall
- Cragg, J. G. and Uhler, R. S. (1970). The demand for automobiles. *Canadian Journal of Economics*, 3, 386-406.
- Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their application*. UK: Cambridge Press.
- Drasgow, F. (1986). Polychoric and polyserial correlations. *The Encyclopedia of Statistics*, 7, 68-74.
- Efron, B. (1978). Regression and ANOVA with zero-one data: measures of residual variation. *Journal of the American Statistical Association*, 73, 113-121.
- Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association* 82, 171-185.
- Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. NY: Chapman and Hall.
- Green, P.E., Carroll, J.D., and DeSarbo, W.S. (1978). A new measure of predictor variable importance in multiple regression. *Journal of Marketing Research*, 15, 356-360.
- Healy, M.J.R. (1990). Measuring importance. *Statistics in Medicine*, 9, 633-637.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12, 55-67.
- Hosmer, D.W., Jovanovic, B. and Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics*, 45, 1265-1270.
- Hosmer, D. W., and Lemeshow, S. (2000). *Applied logistic regression*. Second Edition, NY: John Wiley & Sons, Inc.
- Jog, V., Zhu, P. and Dutta, S. (2006). One share-one vote. *Canadian Investment Review*, Fall, 9-13.
- Kruskall, W. (1987). Relative importance by averaging over orderings. *The American Statistician*, 41, 6-10.
- Laitila, T. (1993). A pseudo-R² measure for limited and qualitative dependent variable models. *Journal of Econometrics*, 56, 341-356.

EXPLANATORY VARIABLES IN A LOGISTIC REGRESSION

McCullagh, P. and Nelder, J.A. (1989). *Generalized linear models*, Second Edition. NY: Chapman and Hall.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. NY: Academic Press, p. 105-142.

McKelvey, R. D. and Zavoina (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103-120.

Mittlbock, M. and Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, 15, 1987-1997.

Nagelkerke, N. J. D. (1991). A note on general definition of the coefficient of determination. *Biometrika*, 78, 691-692.

Nordberg, L. (1981) Stepwise selection of explanatory variables in the binary logit model. *Scandinavian Journal of Statistics*, 8, 17-26.

Platt, R. W. Hanley, J. A., and Yang, H (2000). Bootstrap confidence intervals for the sensitivity of a quantitative diagnostic test. *Statistics in Medicine*, 19, 313-322.

Pratt, J.W. (1987). Dividing the indivisible: using simple symmetry to partition variance explained. In T. Pukkila and S. Puntanen (Eds.), *Proceedings of the Second International Conference in Statistics* (p. 245-260). Tampere, Finland: University of Tampere.

Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705-724.

Thomas, D. R. (1992). Interpreting discriminant functions: a data analytic approach. *Multivariate Behavioral Research*, 27, 335-362.

Thomas, D.R., Hughes, C.E. and Zumbo, B.D. (1998). On variable importance in linear regression. *Social Indicators Research: An International and Interdisciplinary Journal for Quality-of-Life Measurement*, 45, 253-275.

Thomas, D. R. and Zumbo, B. D. (1996). Using a measure of variable importance to investigate the standardization of discriminant coefficients. *Journal of Educational and Behavioral Statistics*, 21, 110-130.

Thomas, D. R. Zhu, P. and Decady Y. (2007). Point estimates and confidence intervals for variable importance in multiple linear regression. *Journal of Educational and Behavioral Statistics*, 32, 61-91.

Windmeijer, F.A.G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Review*, 14, 101-116.

Second-Order Latent Growth Models with Shifting Indicators

Gregory R. Hancock
University of Maryland

Michelle M. Buehl
George Mason University

Second-order latent growth models assess longitudinal change in a latent construct, typically employing identical manifest variables as indicators across time. However, the same indicators may be unavailable and/or inappropriate for all time points. This article details methods for second-order growth models in which constructs' indicators shift over time.

Key words: latent growth modeling; structural equation modeling; curve-of-factors models.

Introduction

Applying structural equation modeling (SEM) techniques to the study of change has become a particularly powerful method for analyzing change over time. Specifically, a special parameterization of SEM called latent curve analysis or latent growth modeling (LGM) has proven to be an extremely flexible approach to study a variety of growth and change questions. LGM provides estimates of many substantively important aspects of change, such as the status of individuals at some substantively interesting temporal reference (e.g., initial measurement point), their growth or change trajectory over time, and the amount of individual variability at a reference point and in rates of growth.

Although techniques such as hierarchical linear modeling can offer such information, LGM also allows one to examine latent correlates or latent predictors of these

growth parameters. Thus, one may not only estimate the form and nature of intra- and inter-individual growth over time, but also test the contribution of other constructs to these growth processes. Further, the flexibility offered by LGM allows for the testing of diverse error structures (e.g., auto-regressive), allows means and variances to change over time, and thus provides a unified assessment of many aspects of change. The basic principles and applications of these models are discussed in several useful didactic sources (e.g., Bollen & Curran, 2005; Byrne & Crombie, 2003; Duncan, Duncan, & Strycker, 2006; Hancock & Lawrence, 2006; Preacher, Wichman, MacCallum, & Briggs, 2008).

Whereas traditional growth models evaluate change in a single measured variable over time, a more complex parameterization evaluates growth in a single unmeasured latent variable (i.e., factor or construct) over time, where that factor has the same multiple measured indicators at each time point. For example, if a child development researcher gathered data using the same five childhood aggression scales at multiple time periods, one could specify relations from, say, an initial status factor and a linear growth factor to each of the latent constructs at each time period. This model has the benefit of analyzing growth using latent constructs disattenuated from measurement error, error that would be present when analyzing only one of the repeated manifest scale values or even some aggregate across scales. Such a model has been referred to as a "curve-of-factors model" (McArdle, 1988), a

Gregory R. Hancock is Professor and Chair in the Department of Measurement, Statistic and Evaluation. His interests include structured means models, latent variable models, and latent growth models. Email: ghancock@umd.edu. Michelle M. Buehl is an Associate Professor of Educational Psychology. Her interests are in student and teacher beliefs, learning, motivation, and academic development. Email her at mbuehl@gmu.edu

“latent variable longitudinal curve model” (Tisak & Meredith, 1990), and, because there are two levels of latent constructs but only one level of manifest indicators, a “second-order latent growth model” (Hancock, Kuo, & Lawrence, 2001; Sayer & Cumsille, 2001).

To elaborate briefly, for change being assessed across T time points, let η_t be a latent construct indicated at time t by J measured variables Y_{tj} ($j=1, \dots, J$). That is, $\mathbf{y} = \boldsymbol{\tau} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\epsilon}$, where the vector $\mathbf{y}$ contains T sets of values across time for J Y variables, $\boldsymbol{\tau}$ is a vector of variable intercepts, $\boldsymbol{\Lambda}$ is a matrix of loadings relating each η_t construct to its measured variable indicators, $\boldsymbol{\eta}$ is a vector of the η_t constructs, and $\boldsymbol{\epsilon}$ is a vector of random normal errors. So far, this is simply a conventional first-order confirmatory factor model with its mean structure modeled simultaneously.

As for modeling growth in the η_t constructs, it can be described by $\boldsymbol{\eta} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$, where $\boldsymbol{\Gamma}$ is a matrix of second-order factor loadings reflecting the hypothesized growth pattern underlying the η_t constructs, $\boldsymbol{\xi}$ is a vector of exogenous latent factors capturing the facets of growth being modeled, and $\boldsymbol{\zeta}$ is a vector of random normal disturbances in the first-order η_t constructs. As an example, a model could posit that $\boldsymbol{\xi} = [\alpha \ \beta]'$, where α is an intercept factor representing the true initial amount of η and where β is a slope factor representing the true rate of linear change in η over time. If the indicators of this η construct are measured at four equal-interval time points, a test for linear growth in the construct could be conducted by fitting second-order factor loadings $\boldsymbol{\Gamma}$ with $[1 \ 1 \ 1 \ 1]'$ in its first column and $[0 \ 1 \ 2 \ 3]'$ in its second. Of course, nonlinear models may be accommodated in this framework as well, as can unequally spaced times points, precisely paralleling first-order growth models.

The illustrative second-order latent growth model described above, referred to herein as Model 1, is depicted in Figure 1. Indicator variables A through E are measured at each of four equal-interval time points. First-order loadings λ are constrained equal across constructs, as are first-order intercepts τ . As

explained elsewhere (e.g., Hancock et al., 2001), indicator variables used to identify factor scales do not require intercepts. Potential variations in this model include the imposition of error variance constraints on the same variable over time, disturbance variance constraints, error covariances for corresponding variables over time (with or without equality constraints), unequally spaced time intervals, alternative temporal reference points, and the inclusion of nonlinear growth constructs. There are five key parameters in this model as specified: κ_α and κ_β , the means of the intercept and slope factors, respectively; ψ_α and ψ_β , the variances of the intercept and slope factors (through their disturbances), respectively; and $\psi_{\alpha\beta}$, the covariance between the intercept and slope factors. This type of model serves as the foundation for a case in which the same variables are not used across all time points, as detailed next. [Note that in Figure 1, as well as Figures 2, 4, and 5, there appear two pseudovariables (unit constants, depicted as a 1 in a triangle). Although common notation utilizes a single such symbol in a given path diagram, two are used here to reduce clutter in each figure.]

Shifting indicators

In all treatments of second-order growth models (e.g., Hancock et al., 2001; McArdle, 1988; Sayer & Cumsille, 2001; Tisak & Meredith, 1990), the assumption is that identical manifest indicators are available at each time period. However, for a variety of phenomena in the social sciences, such an assumption may be unreasonable (see, e.g., McArdle, 1994). For example, when assessing fear responses from infancy through early childhood to ascertain if children become more or less fearful over time, as well as the determinants of such development, the process of eliciting and measuring a fear response must differ at different ages.

Quite simply, some stimuli that frighten a 6-month old might not frighten the child when reaching 12 or 18 months of age, and children might not demonstrate fear in the same manner over time as their ability to communicate develops.

Figure 1. Standard second-order latent growth model ("Model 1")

Figure 2. Second-order latent growth model with shifting indicators ("Model 2")

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

Similarly, when administering survey items to children over a time period before and after becoming literate, rating scales may start with oral administration by a teacher, follow later by having the child circle “smiley face” responses, and ultimately end by having the child circle verbal descriptors. As a final example, an organization assessing employee satisfaction longitudinally might find that the wording of some of the items falls out of common usage. This was the case with the very popular Job Descriptive Index (JDI), where several satisfaction items were changed because the wording of some items was no longer in the popular vocabulary (Smith et al., 1987). Thus, reminiscent of issues of equating in item response theory (see, e.g., Kolen & Brennan, 2004), what is needed here is a way of analyzing growth when the latent construct is conceptualized to be the same over time but the manifest indicators shift or change.

Consider the developmental researcher who wishes to investigate growth in a latent construct across equally-spaced time points in children’s lives, and ideally would like to be able to obtain the same five measurements at each time as depicted in Figure 1 (i.e., measures A, B, C, D, and E). Thus, the same five indicator variables would be used at each time point in a second-order latent growth model, with corresponding first-order loadings and intercepts constrained equal over time. Imagine, however, that at Time 1 only measures A and B are developmentally appropriate; that is, the child is not yet ready to face the stimuli or tasks required for measures C, D, and E. Further, at Time 2, measure A is now too simple and only measures B and C are developmentally appropriate; D and E are still too advanced; and so forth. So, the indicators are shifting as required by developmental considerations, and the actual set of available indicator variables is thus as follows:

<u>Time 1</u>	<u>Time 2</u>	<u>Time 3</u>	<u>Time 4</u>
A1			
B1	B2		
	C2	C3	
		D3	D4
			E4

Notice first that Time 1 and Time 3 share no common indicators, nor do Time 2 and Time 4. However, there is overlap between adjacent time points, such that each construct is linked to its temporal neighbor through a common indicator variable. Linking among constructs, of which many forms exist (as discussed below), will be necessary to facilitate model identification for a second-order growth model with shifting indicators. The model just described, referred to herein as Model 2, is depicted in Figure 2 above.

In addition to the lack of any common indicators across all time points, many features are noteworthy about Model 2. First, and as before, corresponding loadings and intercepts are constrained equal across time; such invariance is crucial for the shifting indicator model to function properly. In the full model, Model 1, variable A was chosen as the scale indicator for all factors. In Model 2 where indicators shift, even though only the first construct has variable A as an indicator, it is still, in fact, the scale indicator for all factors by virtue of the loading constraints across factors. A one unit increase in η_1 yields a one unit increase in A (A1) and a λ_B unit increase in B (B1); a one unit increase in η_2 also yields a λ_B unit increase in B (B2), as well as a λ_C unit increase in C (C2); a one unit increase in η_3 also yields a λ_C unit increase in C (C3), as well as a λ_D unit increase in D (D3); a one unit increase in η_4 also yields a λ_D unit increase in D (D4), as well as a λ_E unit increase in E (E5).

Thus, all variables are linked back to the first construct through equality constraints on λ_B , λ_C , and λ_D , and to the units assigned the first construct through its scale indicator. Note also that any other variable could have been chosen as the scale indicator, resulting only in a change of the metric of parameters’ solutions. If variable B had been selected, for example, it would have a unit loading on both η_1 and η_2 , thereby constraining those loadings implicitly; the C and D loadings would be constrained explicitly with formal equality constraints, while the A and E loadings would be free. Whether A or B is chosen as the scale indicator, or any other variable for that matter, only four unique loadings are estimated.

Second, with regard to intercepts, corresponding parameters are constrained over time just as in Model 1, and variables assigned the role of scale indicator have no intercept term estimated. In Model 2, just as in the full model, only four intercepts will be estimated: one for each variable other than the system's scale indicator.

Third, just as in Model 1, there are five key parameters to be estimated: κ_α and κ_β , the means of the intercept and slope factors, respectively; ψ_α and ψ_β , the variances of the intercept and slope factors (through their disturbances), respectively; and $\psi_{\alpha\beta}$, the covariance between the intercept and slope factors. As will be illustrated below, the parameters estimated under this reduced model with shifting indicators are the same as those under the full model.

Finally, many variations and extensions to this model are possible. Error variances for the same variable may be constrained over time, reflecting comparable measurement error in each variable across time. Similarly, disturbance variances in the first order constructs may be constrained equal. Also, error covariance parameters may be estimated for corresponding variables over time (with or without equality constraints). Unequally spaced time intervals may be accommodated, alternate temporal reference points may be employed, and nonlinear growth constructs may be included under similar configurations. And lastly, as discussed below, many configurations of linked variables could make such a model identified.

Configuration requirements for shifting indicators

As mentioned earlier, many practical reasons might give rise to a shifting set of indicator variables. Developmental necessities, for example, could yield a pattern as seen in Model 2. Administrative decisions within a company, on the other hand, could change the content of the evaluation instruments as new issues arise. In fact, one could simply encounter a loss or corruption of data gathered at different points in time, or errors in measure administration could result in some items being mistakenly omitted from a survey. Regardless of the mechanism giving rise to the particular

shifting pattern of the indicators, all factors must be linked to each other through loading constraints (equality or unit scaling) either directly or indirectly; otherwise a consistent metric for constructs is not preserved over time.

In most practical situations, one could most likely inspect the available indicators at each time point to see if adequate construct linking exists. In more complex longitudinal systems, however, establishing the necessary linking might be less clear by inspection alone. A heuristic is thus offered for establishing this sufficient condition. Consider a first-order factor repeated over T time points, with J_t indicators at each time. All first-order factor loadings are in

matrix $\mathbf{\Lambda}$, which has $\sum_{t=1}^T J_t$ rows and T columns.

Figure 3 depicts the loading matrix for Model 2 from Figure 2. The information in $\mathbf{\Lambda}$ may be abbreviated in a symbolic $p \times T$ configuration matrix $\mathbf{C}$, where p is the number of unique variables across all T time points (i.e., the number of distinct elements in the union of the T indicator sets, each of which has J_t elements). The 5×4 configuration matrix $\mathbf{C}$ for Model 2 is also shown in Figure 3, where asterisks indicate a variable loading on a construct at one or more time points. Next, from $\mathbf{C}$ a $T \times T$ incidence matrix $\mathbf{M}$ may be derived such that, for $t=1$ to T and $u=1$ to T , element $m_{tu}=1$ if the t th and u th constructs have one or more common constrained indicator variables (measured at times t and u) and element $m_{tu}=0$ otherwise. If $t=u$, then obviously $m_{tu}=1$. The incidence matrix $\mathbf{M}$ for Model 2 appears in Figure 3. Finally, drawing from Markov-chain treatments of the decomposition of a state-space into equivalence classes (see, e.g., Ross, 2000), a heuristic for assessing whether sufficient linkages exist among constructs comes by assessing whether matrix $\mathbf{M}$ is irreducible. Specifically, if raising $\mathbf{M}$ to the T th power yields a matrix with all positive elements, then the matrix is irreducible and sufficient linkages exist. If, however, any zero elements are present in $\mathbf{M}^T$, then the constructs are not all linked, and a second-order growth model with shifting indicators cannot be fit to the data. Because elements e_{tu} in $\mathbf{M}^T$ are all nonnegative, this operationalizes sufficient

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

linkage as $\sum_{t=1}^T \sum_{u=1}^T e_{tu} > 0$. Although in the case

of Model 2 inspection alone is enough to establish sufficient construct linkage, as shown in Figure 3 the resulting matrix $\mathbf{M}^T$ contains all positive elements.

$$\Lambda = \begin{array}{c} \begin{array}{ccccc} & \eta_1 & \eta_2 & \eta_3 & \eta_4 \\ \begin{array}{c} A1 \\ B1 \\ B2 \\ C2 \\ C3 \\ D3 \\ D4 \\ E4 \end{array} & \begin{bmatrix} 1 & 0 & 0 & 0 \\ \lambda_B & 0 & 0 & 0 \\ 0 & \lambda_B & 0 & 0 \\ 0 & \lambda_C & 0 & 0 \\ 0 & 0 & \lambda_C & 0 \\ 0 & 0 & \lambda_D & 0 \\ 0 & 0 & 0 & \lambda_D \\ 0 & 0 & 0 & \lambda_E \end{bmatrix} \end{array} \end{array}$$

$$\mathbf{M} = \begin{array}{c} \begin{array}{ccccc} & \eta_1 & \eta_2 & \eta_3 & \eta_4 \\ \begin{array}{c} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{array} & \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \end{array} \end{array}$$

$$\mathbf{M}^T = \begin{array}{c} \begin{array}{ccccc} & \eta_1 & \eta_2 & \eta_3 & \eta_4 \\ \begin{array}{c} \eta_1 \\ \eta_2 \\ \eta_3 \\ \eta_4 \end{array} & \begin{bmatrix} 9 & 12 & 9 & 4 \\ 12 & 18 & 16 & 9 \\ 9 & 16 & 18 & 12 \\ 4 & 9 & 12 & 9 \end{bmatrix} \end{array} \end{array}$$

Figure 3. Matrices associated with Model 2

It should be also noted that the above criterion of incidence matrix irreducibility does not actually constitute a minimum condition for model identification. Regardless of how many indicators are present at each of the T time points, the minimum condition for model identification requires that $T-1$ pairwise constraints (equality or unit scaling) must exist and in a specific configuration. This can be operationalized using the elements below the diagonal of the incidence matrix $\mathbf{M}$, requiring that a minimum of $T-1$ nonzero elements be arranged such that every combination of t th row and u th column has at least one nonzero lower triangular element in its union (note that multiple such configurations exist). Put simply, if one draws horizontal and vertical lines through and beyond each of the $T-1$ nonzero

entries below the diagonal of $\mathbf{M}$, minimum identification conditions have been met if and only if every lower-triangular cell has at least one line passing through it. This heuristic of course works with more than the minimum $T-1$ constraints as well.

Examples

Two examples are offered in this section. The first is for a population matrix in which five indicators are present at each of four time points. This model will be analyzed in a full second-order latent growth model form as in Model 1, and a minimal reduced form as in Model 2, showing key parameters to be equivalent in both solutions. The second example will draw from the National Education Longitudinal Survey of 1988 (NELS:88) data set, sponsored by the National Center for Educational Statistics, U.S. Department of Education (see Ingels, Dowd, Baldrige, Stipe, Bartot, & Frankel, 1994). In this example, sample data for four indicators are present at each of three time points. The model will be analyzed in full second-order latent growth model form, and then all possible minimal reduced forms (using the same scale indicator). Summary information for all reduced forms will be presented and compared to the results from the full sample.

Example 1: Contrived Population Data

Population data were created for all 20 variables in Model 1, where the same five indicators are used for a factor at each of four equally-spaced time points. These data, which consist of a 20×20 population covariance matrix and 20 population means, are embedded within the EQS 6.1 syntax (Bentler, 2004) for this example presented in Appendix A. [This model could be run in any standard SEM software; EQS was chosen merely for illustration as seen in Appendices A and B.] Note that in this program the sample size for this population was arbitrarily set to 100,000; this choice does not affect parameter estimation.

A full second-order latent growth model was imposed upon the data as shown in Figure 1 (with intercept factor α , linear growth factor β , and loading and intercept constraints), and allowed error covariances between residuals of

common indicators at multiple time points (e.g., ϵ_{A1} , ϵ_{A2} , ϵ_{A3} , ϵ_{A4}). No variance or covariance equality constraints were imposed. This model resulted in perfect data-model fit (given the contrived nature of the data) and yielded the following (expected) key parameter solutions: $\kappa_{\alpha}=1.400$, $\kappa_{\beta}=0.700$, $\psi_{\alpha}=0.490$, $\psi_{\beta}=0.490$, $\psi_{\alpha\beta}=-0.098$.

Now imagine having population data for only the eight indicators in the shifting indicator model as depicted in Figure 2. These data, which consist of an 8×8 population covariance matrix and eight population means, are embedded within the EQS 6.1 syntax (Bentler, 2004) for this example also presented in Appendix A. Again the sample size for this population was arbitrarily set to 100,000.

A second-order latent growth model with shifting indicators was imposed as shown in Figure 2 (intercept factor α , linear growth factor β , intercept and loading constraints). As before the A variable loading is set to unity, but in this case there is only one A variable, that at the first time point (i.e., A1). Even though this model does not have the same indicator variables directly present across factors, and even though there is no loading fixed to unity for the construct at times 2, 3, and 4, a solution emerges that is identical to that from the full set: $\kappa_{\alpha}=1.400$, $\kappa_{\beta}=0.700$, $\psi_{\alpha}=0.490$, $\psi_{\beta}=0.490$, $\psi_{\alpha\beta}=-0.098$. As discussed previously, this phenomenon arises because the imposition of constraints effectively forces the first factor's scale indicator A1 to be the scaling indicator for all constructs even though variable A does not load on them. Further, any set of indicators meeting the configuration criteria previously described will yield identical parameter solutions as long as variable A provides scale directly and indirectly to all first-order factors. Thus, the shifting indicator model can capture the parameter estimates using only a subset of the indicators.

Example 2: NELS Data for Full and Reduced Second-Order Growth Models

As a second illustration of second-order growth models with shifting indicators, data from 228 females from the NELS:88 data set

were used. Specifically, the construct of self-concept was assessed at 8th, 10th, and 12th grades, using a variety of indicator variables. Four indicator variables were selected that seemed most theoretically related to students' self-concept. These items were based on Rosenberg's (1965) widely used measure of self-esteem and included: "On the whole, I feel good about myself;" "I feel I am a person of worth;" "On the whole, I feel satisfied with myself;" and "At times, I think I am no good at all." Respectively, these are items 44A, 44D, 44H, and 44J from 8th grade, 62A, 62D, 62H, and 62J from 10th grade, and 66A, 66D, 66H, and 66J from 12th grade. All measures utilized a four-point Likert format; for the current example all responses were recoded such that a higher numerical response on a variable represented a more positive self-concept. Although a compelling argument could be made for treating these data as ordinal, we will treat them as intervally scaled measures for the purposes of illustration. Summary statistics for these data are embedded within the EQS 6.1 syntax (Bentler, 2004) for this example, presented in Appendix B.

First, as a frame of reference, a second-order latent growth model was fit to these data as shown in Figure 4, with intercept factor α and linear growth factor β , assuming equally-spaced time points, constraining first-order loadings and intercepts to be equivalent across shared (adjacent) time points, constraining error variances for common variables and first-order disturbance variances to be equal over time, allowing nonzero error covariances for common indicator variables over time (not shown in figure), and using variable J (i.e., 44J, 62J, 66J) as the first-order factors' scale indicator.

The comparative fit index (CFI) and root mean-square error of approximation (RMSEA), as well as key parameter estimates for the full model, appear in Table 1. By even the most modern and rigorous of standards (e.g., Hu & Bentler, 1999), this data-model fit was excellent. The parameter estimates indicate interesting

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

Figure 5. Example of second-order latent growth model with shifting indicators, for NELS:88 data

intercept behavior, but rather uninteresting slope behavior (small mean and variance, neither statistically significant). This is probably due to the very stable nature of global self-esteem assessed by the Rosenberg (1965) measure.

Second, using the same NELS:88 data for 228 females, a complete set of shifting indicator models was conducted with the following characteristics: only two indicators were present at each time point, no variable appeared at more than two time points, adjacent time points were linked with a single common indicator variable, and variable J was present in all models as the scale indicator (appearing either at one or two time points). As with the full model, each of these minimal shifting indicator models had corresponding loadings, intercepts, and error variances constrained, as well as error covariances between common indicators' residuals at adjacent time points and constrained first-order disturbance variances. A total of 24 such configurations existed and were run on these data; an example of such a model is depicted in Figure 5. Summaries of data-model fit as well as means and medians of key parameter estimates are presented in Table 1.

Whereas the previous example illustrated that population values will be identical for the full and reduced models, sample values will vary across reduced models. This is because the full model imposes constraints across factors at all time points (i.e., their indicators' loadings, intercepts, and error variances), but in these reduced models such constraints exist only in temporally adjacent factors. Thus, in applied scenarios when data exist with only select indicators available at each time point, and in a shifting but linked configuration, one can expect results to be somewhat dependent upon the variables at hand. Still, in the current example when averaging across all of the reduced models, the typical inferences regarding each of the key parameter estimates do match those of the full model. In general, such coherence will be expected to be enhanced the more the model at hand, and all its constraints, constitute a sound approximation to the true growth process operating in the population, as is true in any latent variable model.

Discussion

The methods presented in the current article have roots in several related areas of modeling. Certainly the principles of latent growth modeling for measured variables, and in particular their second-order adaptation for growth in latent variables, are foundational. Also related are growth modeling methods for accelerated longitudinal designs with measured (or latent) variables (see, e.g., Duncan, Duncan, & Hops, 1996; McArdle, 1994; McArdle & Bell, 2000; McArdle & Hamagami, 1991; McArdle & Woodcock, 1997). In such designs interest still resides in gaining an understanding of development over T time points, but specifically in doing so without following the same group of individuals for the entire period. Rather, concurrent cohorts of individuals with adjacent and overlapping subsets of time points (e.g., Cohort 1 at ages 4, 5, and 6, Cohort 2 at ages 6, 7, and 8, and Cohort 3 at ages 8, 9, and 10) are essentially spliced together through constraints on common parameters within a multisample latent growth model. The current work also effectively splices together parts of a model by constraining common parameters, but does so all within a single sample followed for the entire T time points. The need for the current method's splicing arises because one is faced with staggered subsets of ("shifting") indicator variables perhaps due to indicators' unavailability or their developmental inappropriateness. As such constructs' common parameters across different time points are constrained in an attempt to give constructs a common identity and thus be able to model growth therein. Minimum constraint conditions involving the incidence matrix $\mathbf{M}$ were presented in this article (and which can, in fact, easily be adapted for accelerated longitudinal designs).

The need for parameter constraints and a common identity for factors also brings up the larger (and much thornier) issue of factorial invariance, both from theoretical and statistical perspectives. First, addressing the theoretical, the second-order latent growth model with shifting indicators is predicated upon the assumptions that (1) the same unidimensional construct exists at all time points, and (2) that

Table 1

Parameter Estimates for Full and Reduced NELS:88 Growth Models

	Full model estimate	Shifting model mean estimate	Shifting model median estimate
CFI	.969	.987	1.000
RMSEA	.039	.015	.000
κ_{α}	2.746*	2.730*	2.671*
κ_{β}	.011	.028	.027
ψ_{α}	.044*	.094*	.077*
ψ_{β}	.006	.017	.012
$\psi_{\alpha\beta}$	.002	-.018	-.008
* $p < .05$			

are adequate to draw inference regarding said construct. If a researcher believes that, for example, the nature of fear in children transforms with age, or that the new survey items do not quite reflect the same construct as those used previously, then the techniques illustrated here should not be used.

Second, regarding the statistical issues of factorial invariance, one might wonder what level of invariance is necessary to give us confidence that the same construct does indeed exist at all time points. Should all common loadings, intercepts and error variances be constrained equal, representing *strict factorial invariance*? Should just common loadings and intercepts be constrained, representing *strong factorial invariance* (Meredith, 1993)? May loadings alone be constrained, representing *weak factorial invariance* (Widaman & Reise, 1997)? May only some loadings and intercepts be constrained, yielding *partial loading invariance* and *partial intercept invariance* (Byrne, Shavelson, & Muthén, 1989)? Such questions are, first, not unique to the growth models at hand, but in fact pervade many model types whether longitudinal or multisample. Second, while discussion of such types of invariance is in no short supply (for a nice didactic treatment see Widaman and Reise, 1997), recommendations

within the methodological and applied literature regarding sufficient invariance conditions to ensure valid structure inference are without general consensus. Nor is it the purpose of this article to attempt to facilitate such consensus, either in the context of the current models or otherwise.

As is often practiced in other scenarios, for assessing second-order latent growth with shifting indicators a family of models representing different degrees of invariance may be tested, ranging from strict factorial invariance (involving all available common parameters) to a model meeting only minimum identifying constraints. Certainly if a model is selected whose constraints are inconsistent with truth, then the ability to make accurate population inferences regarding growth could become compromised. On the other hand, if a model is selected whose constraints perfectly mirror the population invariance (whether strict, partial, or in between), growth in the construct can indeed be modeled as demonstrated in the current article as long as minimum identification conditions are met. [It should be noted that these minimum conditions, while sufficient for model identification, actually render the structural parameters of interest locally just-identified; having additional common indicators across

time (i.e., two or more per factor) therefore allow for improved structural assessment in terms of both testability and parameter estimation.] However, even satisfactory data-model fit for the strictest of invariance conditions cannot guarantee that the nature of the construct is intransient, but rather can only lend support to the theory of stability of the construct's identity. That is to say, ultimately the stability of a construct's identity rests with strong theoretical foundations regarding the construct as well as the indicators themselves, for which tests of invariance may provide confirmatory evidence. For cases where a stably defined construct is hypothesized, it is expected that the methods illustrated in the current article (and the many variations thereof) will have wide-ranging applications in scenarios where developmental and/or administrative conditions have dictated the absence of common variables across all time points.

References

- Bentler, P.M. (2004). *EQS structural equations program*. Encino, CA: Multivariate Software, Inc.
- Bollen, K. A., & Curran, P. J. (2005). *Latent curve models: A structural equation perspective*. NY: Wiley.
- Byrne, B. M., & Crombie, G. (2003). Modeling and testing change: An introduction to the latent growth curve model. *Understanding Statistics*, 2, 177-203.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105, 456-466.
- Duncan, S. C., Duncan, T. E., & Hops, H. (1996). Analysis of longitudinal data within accelerated longitudinal designs. *Psychological Methods*, 1, 236-248.
- Duncan, T. E., Duncan, S. C., & Strycker, L. A. (2006). *An introduction to latent variable growth curve modeling: Concepts, Issues, and Applications* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hancock, G. R., Kuo, W-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, 8, 470-489.
- Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (p. 171-196). Greenwood, CT: Information Age Publishing, Inc.
- 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, 1-55.
- Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H., & Frankel, M. R. (1994). *National education longitudinal study of 1988, second follow-up: Student component data file user's manual* (Report No. NCES 94-374). Washington, DC: U.S. Dept. of Education, Office of Educational Research and Improvement.
- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). NY: Springer.
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., p. 561-614). New York: Plenum Press.
- McArdle, J. J. (1994). Structural factor analysis experiments with incomplete data. *Multivariate Behavioral Research*, 29, 409-454.
- McArdle, J. J. & Bell, R. Q. (2000). Recent trends in modeling longitudinal data by latent growth curve methods. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and scientific examples* (p.69-107). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- McArdle, J. J., & Hamagami, F. (1991). Modeling incomplete longitudinal and cross-sectional data using latent growth structural equation models. In L. M. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (p. 276-304). Washington, DC: American Psychological Association.

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

McArdle, J. J., & Woodcock, R. W. (1997). Expanding test-retest designs to include developmental time-lag components. *Psychological Methods*, 2, 403-435.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.

Preacher, K. J., Wichman, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Thousand Oaks, CA: Sage.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Ross, S. M. (2000). *Introduction to probability models* (7th ed.). Burlington, MA: Harcourt Academic Press.

Sayer, A.G. & Cumsille, P. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (p. 177-200). Washington, DC: American Psychological Association.

Smith, P. C., Balzer, W., Brannick, M., Chia, W., Eggleston, S., Gibson, W., et al. (1987). The revised JDI: A facelift for an old friend. *The Industrial-Organizational Psychologist*, 24, 31-33.

Tisak, J., & Meredith, W. (1990). Descriptive and associative development models. In A. Von Eye (Ed.), *Statistical methods in longitudinal research* (Vol. 2, p. 387-406). Boston, MA: Academic Press.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In Kendall, J. B., Windle, M., & West, S. G. (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Acknowledgements

We are most grateful for the thoughtful comments, discussion, and assistance of our colleagues Eric Slud, Paul Smith, Robert Ployhart, Patrick Curran, Jack McArdle, and Kristopher Preacher.

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

Appendix A

EQS syntax for full and reduced population models in Example 1

EQS syntax for Model 1 (full model):

```
/TITLE
Example 1, full model
/SPECIFICATIONS
cases=100,000; variables=20; matrix=cov; method=ml; analysis=moment; fields=10;
/LABELS
V1 = AA1; V2 = BB1; V3 = CC1; V4 = DD1; V5 = EE1;
V6 = AA2; V7 = BB2; V8 = CC2; V9 = DD2; V10 = EE2;
V11 = AA3; V12 = BB3; V13 = CC3; V14 = DD3; V15 = EE3;
V16 = AA4; V17 = BB4; V18 = CC4; V19 = DD4; V20 = EE4;
F1 = ETA1; F2 = ETA2; F3 = ETA3; F4 = ETA4; F5 = ALPHA; F6 = BETA;
/EQUATIONS
V1 = 1F1 + 0V999 + E1;
V2 = *F1 + *V999 + E2;
V3 = *F1 + *V999 + E3;
V4 = *F1 + *V999 + E4;
V5 = *F1 + *V999 + E5;
V6 = 1F2 + 0V999 + E6;
V7 = *F2 + *V999 + E7;
V8 = *F2 + *V999 + E8;
V9 = *F2 + *V999 + E9;
V10 = *F2 + *V999 + E10;
V11 = 1F3 + 0V999 + E11;
V12 = *F3 + *V999 + E12;
V13 = *F3 + *V999 + E13;
V14 = *F3 + *V999 + E14;
V15 = *F3 + *V999 + E15;
V16 = 1F4 + 0V999 + E16;
V17 = *F4 + *V999 + E17;
V18 = *F4 + *V999 + E18;
V19 = *F4 + *V999 + E19;
V20 = *F4 + *V999 + E20;
F1 = 1F5 + 0F6 + D1;
F2 = 1F5 + 1F6 + D2;
F3 = 1F5 + 2F6 + D3;
F4 = 1F5 + 3F6 + D4;
F5 = *V999 + D5;
F6 = *V999 + D6;
/VARIANCES
E1 to E20=*;
D1 to D6=*;
/COVARIANCES
D5,D6 =*;
E1,E6=*; E1,E11=*; E1,E16=*; E6,E11=*; E6,E16=*; E11,E16=*;
E2,E7=*; E2,E12=*; E2,E17=*; E7,E12=*; E7,E17=*; E12,E17=*;
E3,E8=*; E3,E13=*; E3,E18=*; E8,E13=*; E8,E18=*; E13,E18=*;
E4,E9=*; E4,E14=*; E4,E19=*; E9,E14=*; E9,E19=*; E14,E19=*
```

HANCOCK & BUEHL

```

E5,E10=*; E5,E15=*; E5,E20=*; E10,E15=*; E10,E20=*; E15,E20=*;
/CONSTRAINTS
(V2,F1)=(V7,F2)=(V12,F3)=(V17,F4);
(V3,F1)=(V8,F2)=(V13,F3)=(V18,F4);
(V4,F1)=(V9,F2)=(V14,F3)=(V19,F4);
(V5,F1)=(V10,F2)=(V15,F3)=(V20,F4);
(V2,V999)=(V7,V999)=(V12,V999)=(V17,V999);
(V3,V999)=(V8,V999)=(V13,V999)=(V18,V999);
(V4,V999)=(V9,V999)=(V14,V999)=(V19,V999);
(V5,V999)=(V10,V999)=(V15,V999)=(V20,V999);
/MATRIX
1.288
0.630    1.575
0.672    0.720    1.468
0.714    0.765    0.816    1.767
0.756    0.810    0.864    0.918    1.672
0.492    0.420    0.448    0.476    0.504    1.929
0.420    0.650    0.480    0.510    0.540    1.1025    1.88125
0.448    0.480    0.612    0.544    0.576    1.176    1.260    2.244
0.476    0.510    0.544    0.778    0.612    1.2495    1.33875    1.428    2.21725
0.504    0.540    0.576    0.612    0.748    1.323    1.4175    1.512    1.6065    2.601
0.494    0.315    0.336    0.357    0.378    1.376    1.260    1.344    1.428    1.512
0.315    0.4375    0.360    0.3825    0.405    1.260    1.450    1.440    1.530    1.620
0.336    0.360    0.584    0.408    0.432    1.344    1.440    1.736    1.632    1.728
0.357    0.3825    0.408    0.5335    0.459    1.428    1.530    1.632    1.834    1.836
0.378    0.405    0.432    0.459    0.686    1.512    1.620    1.728    1.836    2.144
0.296    0.210    0.224    0.238    0.252    1.668    1.680    1.792    1.904    2.016
0.210    0.425    0.240    0.255    0.270    1.680    2.000    1.920    2.040    2.160
0.224    0.240    0.356    0.272    0.288    1.792    1.920    2.148    2.176    2.304
0.238    0.255    0.272    0.489    0.306    1.904    2.040    2.176    2.512    2.448
0.252    0.270    0.288    0.306    0.424    2.016    2.160    2.304    2.448    2.692
3.150
2.625    3.7125
2.800    3.000    3.900
2.975    3.1875    3.400    4.5125
3.150    3.375    3.600    3.825    4.750
3.140    3.150    3.360    3.570    3.780    5.947
3.150    3.475    3.600    3.825    4.050    5.4075    6.49375
3.360    3.600    4.040    4.080    4.320    5.768    6.180    7.492
3.570    3.825    4.080    4.435    4.590    6.1285    6.56625    7.004    8.14175
3.780    4.050    4.320    4.590    5.060    6.489    6.9525    7.416    7.8795    9.243
/MEANS
1.400    2.500    3.600    4.700    5.800    2.100    3.250    4.400    5.550    6.700
2.800    4.000    5.200    6.400    7.600    3.500    4.750    6.000    7.250    8.500
/PRINT
fit=all;
/END

```

SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

EQS syntax for Model 2 (reduced model):

```

/TITLE
Example 1, Model 2 (reduced model)
/SPECIFICATIONS
cases=100,000; variables=8; matrix=cov; method=ml; analysis=moment; fields=10;
/LABELS
V1 = AA1; V2 = BB1; V3 = BB2; V4 = CC2;
V5 = CC3; V6 = DD3; V7 = DD4; V8 = EE4;
F1 = ETA1; F2 = ETA2; F3 = ETA3; F4 = ETA4; F5 = ALPHA; F6 = BETA;
/EQUATIONS
V1 = 1F1 + 0V999 + E1;
V2 = *F1 + *V999 + E2;
V3 = *F2 + *V999 + E3;
V4 = *F2 + *V999 + E4;
V5 = *F3 + *V999 + E5;
V6 = *F3 + *V999 + E6;
V7 = *F4 + *V999 + E7;
V8 = *F4 + *V999 + E8;
F1 = 1F5 + 0F6 + D1;
F2 = 1F5 + 1F6 + D2;
F3 = 1F5 + 2F6 + D3;
F4 = 1F5 + 3F6 + D4;
F5 = *V999 + D5;
F6 = *V999 + D6;
/VARIANCES
E1 to E8=*;
D1 to D6=*;
/COVARIANCES
D5,D6=*;
E2,E3=*; E4,E5=*; E6,E7=*;
/CONSTRAINTS
(V2,F1)=(V3,F2);
(V4,F2)=(V5,F3);
(V6,F3)=(V7,F4);
(V2,V999)=(V3,V999);
(V4,V999)=(V5,V999);
(V6,V999)=(V7,V999);
/MATRIX
1.288
.630    1.575
.420    .650    1.881
.448    .480    1.260    2.244
.336    .360    1.440    1.736    3.900
.357    .383    1.530    1.632    3.400    4.513
.238    .255    2.040    2.176    4.080    4.435    8.142
.252    .270    2.160    2.304    4.320    4.590    7.880    9.243
/MEANS
1.400    2.500    3.250    4.400    5.200    6.400    7.250    8.500
/PRINT
fit=all;
/END

```

Appendix B

EQS syntax for full NELLS:88 model in Example 2

```

/TITLE
Full model for females' self-concept data
/SPECIFICATIONS
cases=228; variables=12; matrix=cor; method=ml; analysis=moment; fields=6;
/LABELS
V1= 44A; V2=44D; V3=44H; V4=44J;
V5=62A; V6=62D; V7=62H; V8=62J;
V9=66A; V10=66D; V11=66H; V12=66J;
F1=SC8; F2=SC10; F3=SC12; F4=ALPHA; F5=BETA;
/EQUATIONS
V1 = *F1 + *V999 + E1;
V2 = *F1 + *V999 + E2;
V3 = *F1 + *V999 + E3;
V4 = 1F1 + 0V999 + E4;
V5 = *F2 + *V999 + E5;
V6 = *F2 + *V999 + E6;
V7 = *F2 + *V999 + E7;
V8 = 1F2 + 0V999 + E8;
V9 = *F3 + *V999 + E9;
V10 = *F3 + *V999 + E10;
V11 = *F3 + *V999 + E11;
V12 = 1F3 + 0V999 + E12;
F1 = 1F4 + 0F5 + D1;
F2 = 1F4 + 1F5 + D2;
F3 = 1F4 + 2F5 + D3;
F4 = *V999 + D4;
F5 = *V999 + D5;
/VARIANCES
E1 to E12 = *;
D1 to D5 = *;
/COVARIANCES
E1,E5 = *; E1,E9 = *; E5,E9 = *;
E2,E6 = *; E2,E10 = *; E6,E10 = *;
E3,E7 = *; E3,E11 = *; E7,E11 = *;
E4,E8 = *; E4,E12 = *; E8,E12 = *;
D4,D5 = *;
/CONSTRAINTS
(V1,F1)=(V5,F2)=(V9,F3);
(V2,F1)=(V6,F2)=(V10,F3);
(V3,F1)=(V7,F2)=(V11,F3);
(E1,E1)=(E5,E5)=(E9,E9);
(E2,E2)=(E6,E6)=(E10,E10);
(E3,E3)=(E7,E7)=(E11,E11);
(E4,E4)=(E8,E8)=(E12,E12);
(D1,D1)=(D2,D2)=(D3,D3);
(V1,V999)=(V5,V999)=(V9,V999);
(V2,V999)=(V6,V999)=(V10,V999);

```


SECOND-ORDER LATENT GROWTH MODELS WITH SHIFTING INDICATORS

```

(V3,V999)=(V7,V999)=(V11,V999);
/MATRIX
1.000
.347 1.000
.564 .291 1.000
.245 .277 .188 1.000
.375 .209 .223 .187 1.000
.146 .228 .133 .126 .457 1.000
.189 .182 .220 .108 .568 .450
.240 .137 .111 .378 .432 .317
.284 .135 .190 .156 .331 .204
.137 .237 .104 .089 .242 .245
.192 .125 .229 .066 .249 .193
.138 .097 .109 .164 .088 .124
1.000
.311 1.000
.280 .223 1.000
.190 .164 .360 1.000
.354 .168 .595 .479 1.000
.178 .334 .428 .226 .338 1.000
/STANDARD DEVIATIONS
.57 .63 .64 .88 .59 .60
.66 .87 .61 .70 .68 .76
/MEANS
3.20 3.30 3.21 2.67 3.21 3.28
3.08 2.71 3.24 3.32 3.16 2.88
/PRINT
fit=all;
/END

```

Coverage Performance of the Non-Central F -based and Percentile Bootstrap Confidence Intervals for Root Mean Square Standardized Effect Size in One-Way Fixed-Effects ANOVA

Guili Zhang
East Carolina University

James Algina
University of Florida

The coverage performance of the confidence intervals (CIs) for the Root Mean Square Standardized Effect Size (RMSSE) was investigated in a balanced, one-way, fixed-effects, between-subjects ANOVA design. The noncentral F distribution-based and the percentile bootstrap CI construction methods were compared. The results indicated that the coverage probabilities of the CIs for RMSSE were not adequate.

Key words: confidence interval, effect size, ANOVA, root mean square standardized effect size, non-central F -distribution, percentile bootstrap, coverage probability, robustness.

Introduction

Reporting an effect size (ES) in addition to or in place of a hypothesis test has been recommended by some statistical methodologists since as early as the 1960s because ESs are recognized as being more appropriate and more informative (Cohen, 1965, 1994; Cumming & Finch, 2005; Finch et al., 2002; Hays, 1963; Meehl, 1967; Nickerson, 2000; Steiger, 2004; Steiger & Fouladi, 1997). In the last two decades, reporting an ES has become mandatory in some editorial policies (Murphy, 1997; Thompson, 1994) and is strongly recommended for American Psychological Association journals. The Publication Manual of the American Psychological Association (2001) states that it is almost always necessary to include some index

of ES or strength of relationship in the results section of a research paper. The APA Task Force on Statistical Inference (Wilkinson and the Task Force on Statistical Inference, 1999) also supports the report of ESs as well as the obligation of researchers to provide confidence intervals (CI) for all principal outcomes. A CI for an ES is recommended as a superior replacement for significance testing because this CI contains all the information found in the significance tests and vital information not provided by the significance tests about the magnitude of effects and precision of estimates (Cohen, 1994; Cumming & Finch, 2001, 2005). A CI indicates the range of population ESs with which the data are consistent. By contrast, a hypothesis test merely indicates whether the data are consistent with a population ES of zero. Because of the obvious advantages of CIs, advocate on the use of ESs and CIs for ESs is “a rapidly rising tide” (Grissom & Kim, 2005).

Dr. Guili Zhang is Assistant Professor of Research and Evaluation Methodology in the Department of Curriculum and Instruction. Her research interests are in applied statistics. Email her at: zhangg@ecu.edu. Dr. James Algina is Professor of Educational Psychology. His research interests are in psychometric theory and applied statistics. Email him at: algina@ufl.edu.

Effect Size Indices and Confidence Intervals in the Two-Group Case

A large number of ES indices have been developed and proposed (Algina et al., 2005a). For example, the number of commonly used ESs measuring separation of two independent samples alone has almost reached a dozen: Cohen's d (Cohen, 1965), Glass's d , Hedges' g (Hedges & Olkin, 1985), two versions of

Cohens d based on trimmed means and Winsorized variances (d_R^+ suggested by Hogarty & Kromrey, 2001 and d_R suggested by Algina and Keselman 2003b; Algina et al., 2005a), eta squared, omega squared, McGraw and Wong's (1992) common language ES (CL), Cliff's dominance statistic (1993, 1996), Kraemer and Andrews γ_1^* (1982), Wilcox and Muska's W (1999), and Vargha and Delaney's A (2000).

Research investigating the performance of the various ES measures is fairly limited. Hedges and Olkin (1985) suggested that Cohen's d evidenced a small sample bias. Hogarty and Kromrey (2001) compared the performance of nine ES indices when they were used in the context of populations with various levels of nonnormality and variance heterogeneity. The nine indices included Cohen's d , Cliff's dominance statistic, g , γ_1^* , CL , A , d_R^+ , a naïve estimator of W and a .632 bootstrap estimator of W . The results indicated that Cohen's d and Hedges' g showed nontrivial sensitivity to violations of normality and homogeneity of variance, which confirmed the concerns raised about the appropriateness of using these indices as indicators of effects in such populations (Kraemer & Andrews, 1982; Wilcox & Muska, 1999). In addition, d_R^+ evidenced severe bias under small sample conditions. Indices CL , γ_1^* and the naïve estimator of W only appeared to be slightly less sensitive than Cohen's d and Hedges' g , but showed pronounced bias under small sample size condition or nontrivial sensitivity to violations of normality and homogeneity of variance. Cliff's dominance statistic and Vargha and Delaney's A showed better performances in producing relatively unbiased estimates and consistent standard errors.

Hess and Kromrey (2004) investigated the performance of the CIs for Cohen's d and Cliff's dominance statistic constructed by using seven CI construction methods: the normal theory Z band, the percentile bootstrap, the bias corrected bootstrap, the bias corrected and accelerated bootstrap (BCa), pivotal, Studentized pivotal, and the Steiger and Fouladi

interval inversion band. Monte Carlo methods were used to compare CI estimates using random samples generated from populations under known and controlled conditions. Across all of the conditions, all of the CI construction methods provided better coverage probabilities for Cliff's dominance statistic than for Cohen's d , with the exception of the Pivotal Bootstrap method.

Cohen's d and Its Confidence Intervals

In the two-group independent samples case, Cohen's d is probably the most widely accepted ES index for a pairwise contrast on means and it is defined as follows:

$$d = \frac{\bar{Y}_2 - \bar{Y}_1}{S} \quad (1)$$

where $\bar{Y}_j$ is the mean for the j th level ($j = 1, 2$), and S is the square root of the pooled variance. The number of observations in a level is denoted by n_j . Cohen's d estimates:

$$\delta = \frac{\mu_2 - \mu_1}{\sigma} \quad (2)$$

where μ_j is the population mean for the j th ($j=1,2$) level, and σ is the population standard deviation, assumed to be equal for both levels.

Reporting a CI for the ES is important as was well put by Wilkinson et al. (1999), "it is hard to imagine a situation in which reporting a dichotomous reject-accept decision is better than reporting an actual p value or, better still, a confidence interval" (p. 599). Steiger and Fouladi (1997) asserted that "a confidence interval conveys more information, in a more naturally usable form, than a significance test." Interests in the accuracy and usefulness of the ESs have motivated explorations of the usefulness and effectiveness of CIs for ESs (Algina & Keselman, 2003a, 2003b; Bird, 2002; Cumming & Fitch, 2001).

An exact CI for δ can be obtained by using the noncentral t distribution when the sample data are normally distributed, the two population have equal variances, and the scores

are independently distributed (Algina et al., 2005a; Cumming & Fitch, 2001; Johnson & Welch, 1940; Serlin & Lapsley, 1985; Steiger & Fouladi, 1997). This CI is the same CI that Hess & Kromrey (2004) referred to as the Steiger and Fouladi inversion method. In this situation, the noncentral t distribution has two parameters: the degrees of freedom, and the noncentrality parameter λ , which is given by

$$\lambda = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\mu_2 - \mu_1}{\sigma} \right) = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \delta. \quad (3)$$

To find a 95% CI for δ , we first use the noncentral t distribution to find a 95% CI for λ , then multiply the two end points of the interval for λ by $\sqrt{(n_1 + n_2)/n_1 n_2}$ to obtain the two end points of a 95% CI for δ . The lower limit of the 95% CI for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic

$$t = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \left(\frac{\bar{Y}_2 - \bar{Y}_1}{S} \right) \quad (4)$$

is the .975 quantile, and the upper limit of the 95% CI for λ is the noncentrality parameter for the noncentral t distribution in which the calculated t statistic is the .025 quantile of the distribution (Algina et al., 2005a, 2006; Steiger & Fouladi, 1997). Algina and Keselman (2003a) adapted this procedure for the dependent samples case.

As noted previously when the population data are normally distributed, the two population have equal variances, and the scores are independently distributed, the noncentral t distribution-based CI is exact. However, when sampling from nonnormal data, the noncentral t distribution-based CI may not have adequate coverage probability in both the independent samples case (Algina & Keselman, 2003a; Kelley, 2005) and dependent samples case (Algina et al., 2005a). Failure to have adequate coverage probability means, for example, that if a nominal 95% CI for δ is computed, the actual

probability that the CI contains the parameter will be different than .95.

Kelley (2005) compared three methods for constructing a CI around Cohen's ES. Specifically, he evaluated noncentral t distribution-based, the percentile bootstrap, and the BCa CIs through a set of simulation studies that involves three conditions of nonnormality, three cases of sample size, and six cases of population ES. Kelley's study indicated that the noncentral t distribution-based CI has inaccurate coverage probability when data are nonnormal. He concluded that when the assumptions of parametric tests are violated, the integrity of the results based on parametric statistical techniques is suspect. The study by Algina et al. (2006) detected the same problem with the noncentral t distribution-based CI in the dependent samples case. In addition, the results from the Hess and Kromrey (2003, 2004) studies also pointed to the inadequate coverage probability issue with the CIs for Cohen's d .

Results from recent studies indicated that in the two-group case, the bootstrap CI is preferable and should be used instead of the noncentral t distribution-based CI. Kelley (2005) asserted that when the normality assumption is false, a CI constructed with the BCa method is more valid than the noncentral t distribution-based CI. When the normality assumption holds, the BCa method will yield results consistent with the parametric results. Therefore, he recommends the use of the BCa method. Like Kelley, Algina et al. (2006) also found that under many conditions the BCa method worked best, although in some cases of data nonnormality, the BCa method did not control probability coverage. By including a wider range of nonnormality than was investigated by Kelly, they found that the BCa method for setting a CI around the population ES is indeed negatively affected by nonnormality. Additionally, they found that the coverage probability declines as sample size decreases and the population ES increases. It is apparent that even with the nonparametric bootstrap construction methods, problem still persists with CIs for Cohen's δ .

The work reported by Algina and Keselman (2003b), Algina et al. (in press, 2005a), and Kelly (2005) indicated that in both the independent samples and dependent samples

cases, CIs for Cohen's δ may be misleading because of poor coverage probability when data are nonnormal. There is a second problem with using Cohen's δ : although it is intended as a measure of group separation, it is not always an adequate measure of group separation. This shortcoming was pointed out by Wilcox and Keselman (2003), and is due to the fact that δ can be dramatically affected by outliers and long-tailed distributions. Cohen's δ is defined by using the usual population means and variances, both of which are least-square parameters. Least-square parameters are not robust, meaning that a small change in the population distribution can strongly affect the parameters. In particular, the usual population mean and variance can be greatly influenced by the existence of extreme observations (outliers) in a distribution. Slight changes in the population distributions, changes that do not have much effect on the separation of the distributions, can substantially alter the value of δ . Therefore, δ can be a very poor measure of group separation, and can grossly misrepresent the degree to which two distributions differ (Algina et al., 2005b; Wilcox & Keselman, 2003).

Root Mean Square Standardized Effect Size and Its Confidence Intervals

Measures of ES in analysis of variance (ANOVA) are measures of the degree of association between a factor and the dependent variable. When it comes to the one-way, fixed-effects, between-subjects ANOVA case, the available generalized ES measures are, but not limited to, eta squared, omega squared, $d_{\max}$, Cohen's f , and the Mean Square Standardized Effect Size (RMSSE) (Olejnik & Algina, 2003; Steiger & Fouladi, 1997). Eta squared and omega squared are estimates of the degree of association. Eta squared is the proportion of the total sum of squares that is attributed to an effect. It is calculated as the ratio of the effect variance to the total variance. Omega squared is an estimate of the dependent variable variance accounted for by the independent variable in the population for a fixed-effects model. The effect size $d_{\max}$ is an overall ES that is calculated by

utilizing the smallest and the largest means where $d_{\max} = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{S}$ (Cohen, 1988), while

Cohen's f and RMSSE are overall ESs that use all of the means and are measures of the standardized average effect in the population across all of the levels of the independent variable. Among these ES measures, the RMSSE, proposed by Steiger and Fouladi (1997), denoted by f^* in our study, was part of the focus of our investigation. RMSSE is a standardized mean difference measure, a generalization of Cohen's δ , and a variant of Cohen's f .

In a balanced, one-way, between-subjects, fixed-effects design, f^* is defined by Steiger & Fouladi (1997) as follows:

$$f^* = \sqrt{\frac{\sum_{j=1}^J (\mu_j - \mu)^2}{(J-1)\sigma^2}} \quad (5)$$

where μ_j is the mean for the j th level, μ is the grand mean, and σ^2 is the within-level variance, which is assumed to be constant across levels. Recall that Cohen (1969)

$$\text{defined } f = \sqrt{\frac{\sum_{j=1}^J (\mu_j - \mu)^2}{J(n-1)\sigma^2}},$$

so f^* is a variation of Cohen's f .

Consider a one-way, fixed-effects ANOVA with n_j observations in the j th group, and J groups. The F statistic is calculated by using

$$F = \frac{MS_B}{MS_W} \quad (6)$$

where

$$MS_B = \frac{\sum_{j=1}^J n_j (\bar{Y}_j - \bar{Y})^2}{J-1} \quad (7)$$

and

$$MS_W = \frac{\sum_{j=1}^J \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2}{N-J} \quad (8)$$

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

In Equation 7 and 8, Y_{ij} is the i th score in group j , $\bar{Y}_j$ is the sample mean for the group j , and $\bar{Y}$ is the grand mean and is calculated by using

$$\bar{Y} = \frac{\sum_{i=1}^{n_i} \sum_{j=1}^J Y_{ij}}{N} \quad (9)$$

Based on expected mean squares in a balanced design, f^* can be estimated by using

$$\hat{f}^* = \sqrt{\frac{MS_B - MS_W}{nMS_W}} = \sqrt{\frac{1}{n}(F-1)} \quad (10)$$

if $F \geq 1$ and by using $\hat{f}^* = 0$, otherwise. Alternatively, based on the expected value of F under normality f^* can be estimated by using

$$\hat{f}^* = \sqrt{\frac{(N-J-2)}{n(N-J)}(F-1)} \quad (11)$$

if $F \geq 1$ and $\hat{f}^* = 0$ otherwise. Both estimates are very similar but the estimate in Equation 10 was used in our study because it does not require the normality assumption in its derivation.

The CIs for Steiger and Fouladi's f^* can be constructed based on the noncentral F distribution (Steiger and Fouladi, 1997; Steiger, 2004). In a one-way, between-subjects, fixed-effects ANOVA, the F statistic with $J-1$ and $N-J$ degrees of freedom has noncentrality parameter

$$\lambda = \frac{\sum_{j=1}^J n_j (\mu_j - \mu)^2}{\sigma^2} \quad (12)$$

Clearly in a balanced design

$$f^* = \sqrt{\frac{\lambda}{n(J-1)}} \quad (13)$$

To find a $100(1 - \alpha)\%$ (95% in our study) CI for f^* , we first use the noncentral F distribution to find a 95% CI for λ . Once the CI on λ is found, we transform the endpoints of the CI for λ by dividing λ by $(J-1)n$ and then take the square root. The result is an exact CI for f^* in the analysis of variance, when its assumptions are met. The lower limit of the 95% CI for λ is the noncentrality parameter for the noncentral F distribution in which the calculated F statistic is the .975 quantile. The upper limit of the 95% CI for λ is the noncentrality parameter for the noncentral F distribution in which the calculated F statistic is the .025 quantile of the distribution.

Purposes of the Study

Constructing a CI for RMSSE by using the noncentral F distribution is based on the assumption that the data are drawn from normal distributions. If data are not normally distributed, the actual coverage probability of the CI may or may not match the nominal level. A method that may be useful for constructing CI for f^* is the percentile bootstrap (Efron and Tibshirani, 1993). Therefore, the performance of the percentile bootstrap on the construction of CIs for f^* was examined in our current study. The purpose of the study is to investigate the coverage performance of the noncentral F distribution-based and the percentile bootstrap CI for f^* .

Methodology

The noncentral F distribution-based and the percentile bootstrap CIs were implemented for all combinations of the following five factors: (a) five population distributions including the normal distribution and four additional cases from the family of the g and h distributions that are nonnormal (Hoaglin, 1983, Martinez & Iglewicz, 1984); (b) two numbers of levels for treatment groups: $J = 3$ and $J = 6$; (c) three cell sample sizes in each treatment; (d) six values of population RMSSEs; (e) two mean configurations: the equally spaced mean configuration and the one extreme mean

configuration. The nominal confidence level for all intervals investigated was .95 and each condition was replicated 2500 times. The number of bootstrap replications in the bootstrap procedure was 1000.

Conditions

Data for all five distributions were generated from the g and h distributions: (a) $g = h = 0$, the standard normal distribution ($\gamma_1 = \gamma_2 = 0$), where $\gamma_1 = \sqrt{\beta_1}$ and is the skewness, and $\gamma_2 = \beta_2$ and is the kurtosis, (b) $g = .76$ and $h = -.098$, a distribution with the skewness and kurtosis of an exponential distribution ($\gamma_1 = 2$, $\gamma_2 = 3$), (c) $g = 0$ and $h = .225$ ($\gamma_1 = 0$ and $\gamma_2 = 154.84$), (d) $g = h = .225$ ($\gamma_1 = 4.90$ and $\gamma_2 = 4673.80$), and (e) $g = 0$ and $h = .109$ ($\gamma_1 = 0$ and $\gamma_2 = 6$), a distribution with the skewness and kurtosis of a double exponential distribution. The four nonnormal distributions cover a wide range of nonnormality including distributions that are quite strongly nonnormal. Such a selection of distributions allows the researcher to investigate the performances of the CIs under a wide range of data conditions. The goal is to find which procedure or procedures are likely to work well over a wide range of distributions because it is impossible for any one simulation to include every possible distribution that might be encountered in real data or to anticipate what types of distributions are realistic in all of social and behavioral science fields.

The numbers of treatment groups investigated were 3 and 6, which cover the likely range encountered in most research in the social and behavioral sciences. The sample sizes in each treatment included were 20, 35, and 50. Such a range seems fairly typical of sample sizes used in social science research, although clearly does not cover sample sizes found in very small or very large studies.

The treatment group means followed two mean configurations: the equally spaced mean configuration and the one extreme mean configuration, which will allow determination of

whether results tend to generalize over configurations.

Six values of f^* were investigated: 0, .1, .25, .40, .55, and .70. Defining

$$\delta_{\max} = \frac{\mu_{\max} - \mu_{\min}}{\sigma} \quad (14)$$

as Cohen's effect size for the largest and smallest means, under the equally spaced mean configurations, these population f^* values approximately correspond to $\delta_{\max}$ of 0, .2, .5, .8, 1.10, and 1.40, respectively. Under the one extreme mean configuration, these population f^* values roughly correspond to $\delta_{\max}$ of 0, .173, .433, .693, .952, and 1.212. Therefore, a f^* of .0 indicates no effect, .1 a small effect, .25 a medium effect, .40 a large effect, and .55 and .70 very large effects.

The nominal confidence level for all intervals investigated was .95 and each condition was replicated 2500 times, assuring sufficient precision for an adequate initial investigation into the sampling behaviors of the CIs. The number of bootstrap replications in the bootstrap procedure was 1000.

Analyses Conducted

The study was designed to investigate the robustness of the noncentral F distribution-based CIs and the percentile bootstrap CIs for f^* to sampling from nonnormal distributions.

Variables conforming to a g and h distributions are transformations of a standard normal distribution. When g and h are both nonzero,

$$Y = \frac{\exp(gZ) - 1}{g} \exp\left(\frac{hZ^2}{2}\right) \quad (15)$$

where Z is a standard normal variable, and Y is the g and h distributed variable. When g is zero,

$$Y = Z \exp\left(\frac{hZ^2}{2}\right) \quad (16)$$

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Standard normal variables (Z_{ij}) were generated by using RANNOR function in SAS (SAS, 1999). Then the Z_{ij} were converted to the desired g and h distributed random variable by using Equation 15 and 16. To create scores corresponding to the selected values of f^* , it is necessary to linearly transform the g and h distributed variables. Data were generated for three samples and six samples in each replication of each condition by the following steps: First, for the first sample n_1 scores were generated from the appropriate distribution. Then n_2 scores from the same distribution were generated and a constant was added to each score. Thirdly, n_3 scores from the same distribution were generated and a constant was added to each score and so forth until n_J scores from the same distribution were generated and a constant was added to each score. The constants were chosen such that the population RMSSE, f^* would equal to the following values: 0, .1, .25, .40, .55, and .70.

For the equally spaced mean configuration, the addition of the constant was accomplished by using

$$Y_{ij} = X_{ij} + (j-1) \sqrt{\frac{12}{J(J+1)}} f^* \sigma, \quad j = 1, \dots, J. \quad (17)$$

For the configuration with one extreme mean, $Y_{ij} = X_{ij}$ for groups $j = 1, \dots, J-1$. For group J the transformation was

$$Y_{ij} = X_{ij} + \sqrt{J} f^* \sigma. \quad (18)$$

To obtain a $(1-\alpha)\%$ (95% in the current study) CI for f^* , the noncentral F distribution is first used to obtain a 95% CI on λ , the noncentrality parameter of the F distribution. Given an observed F statistic with a value F and known degrees of freedoms, a $(1-\alpha)\%$ CI on λ can be obtained with the following steps (Steiger, 2004):

1. Calculate the cumulative probability of the value F in the central F distribution. This is $1 - p$, where p is the probability level printed by most analysis of variance procedures. If $1 - p$ is below $\alpha/2$, then both limits of the CI are zero. If $1 - p$ is below $1 - \alpha/2$, the lower limit of the CI is zero, and the upper limit must be calculated (go to step 3). Otherwise, calculate both limits of the CI for λ by using steps 2 and 3.
2. To calculate the lower limit of λ , find the unique value of λ that places the F statistic at the $1 - \alpha/2$ probability point of a noncentral F distribution with the known degrees of freedom.
3. To calculate the upper limit of λ , find the unique value of λ that places the F statistic at the $\alpha/2$ cumulative probability point percentile of a noncentral F distribution.

In summary, calculating a CI for λ requires iterative calculation of the unique value of λ that places an observed value of F at a particular percentile of the noncentral F distribution. These procedures were implemented by using the "FNONCT" function in SAS. Notice the CI for f^* constructed by the noncentral F distribution-based method will result in coverage probability of .975 when $f^* = 0$ because the probability noncoverage from the lower side of the distribution will be 0 instead of .025.

Once the CI on λ is found, the endpoints of the CI for λ are transformed to endpoints for f^* by dividing by $(J-1)n$ and then taking the square root. The result is an exact CI for f^* in the analysis of variance, when the ANOVA assumptions are met.

To apply the percentile bootstrap method, the following steps are completed 1000 times within each replication of a condition.

1. A sample of size n_j is randomly selected with replacement from the scores for the group j , $j = 1, \dots, J$.

These J samples are combined to form a bootstrap sample.

2. The parameter f^{*2} is estimated by using

$$\hat{f}^{*2} = \frac{1}{n}(F - 1) \quad (19)$$

1. The 1000 $\hat{f}^{*2}$ estimates are then ranked from low to high. The lower limit of the CI for f^{*2} is determined by finding the 26th estimate in the rank order [i.e., the $(.025 \times 1000 + 1)$ th estimate]; and the 975th estimate is the upper limit of the CI for f^{*2} [i.e., the $(.975 \times 1000)$ th estimate].
2. The lower limit of the CI for f^* is equal to the square root of the lower limit of the CI for f^{*2} if the latter lower limit is larger than zero and is zero otherwise. The upper limit of the CI for f^* is equal to the square root of the upper limit of the CI for f^{*2} .

Results

The estimated coverage probabilities for and the average widths of the noncentral F distribution-based and bootstrap CIs for f^* are reported and compared for all conditions. The estimated coverage probabilities of the noncentral F distribution-based and bootstrap CIs for f^* are reported in Table 1 through Table 4. The average widths of the noncentral F distribution-based and bootstrap CIs for f^* are shown in Table 5 through Table 8.

Estimated Coverage Probabilities of Confidence Intervals for f^*

The interval [.925, .975] used by Algina et al. (2006) was used as a criterion for adequate coverage probability when the nominal confidence coefficient is .95. This interval corresponds to Bradley's (1978) liberal criterion for a nominal .05 Type I error rate. In addition, because this interval may be considered as too lenient, a more stringent interval, [.94, .96], was also used to judge the adequacy of the coverage probabilities. In Tables 1 through 4, estimates that are outside the [.94, .96] interval are bolded,

while estimates that are outside of the interval [.925, .975] are bolded and underlined.

The patterns of results across Tables 1 to 4 for the noncentral F distribution-based CI for f^* are fairly similar. First, when sampling from a normal distribution, as stated earlier, the coverage probability of the noncentral F distribution-based CI should be .975 when $f^* = 0$, and the results in Tables 1 to 4 are consistent with the theory. When $f^* > 0$, the coverage probability of the noncentral F distribution-based CI is expected to be .95 under normality and the results in Tables 1 to 4 are consistent with this expectation.

Second, coverage probability for the noncentral F distribution-based CI tends to be better than for the bootstrap CI both when sampling from normal and nonnormal distributions. When $J = 3$ and samples are drawn from a normal distribution, coverage probability for the noncentral F distribution-based CI is outside [.925, .975] in 2 out of 36 total cases, while the bootstrap CI coverage probability is outside [.925, .975] in 13 cases. Under normality, when $J = 6$, although both CIs have 2 coverage probabilities that are outside [.925, .975], the noncentral F distribution-based CI has 6 coverage probabilities that are outside [.94, .96] while the bootstrap CI has 18 coverage probabilities that are outside this interval. When sampling from the nonnormal distributions, the noncentral F distribution-based CI has fewer coverage probabilities that are outside the criterion intervals than does the bootstrap CI under each of the four distribution conditions.

Third, the performances of the noncentral F distribution-based CIs for f^* under the four nonnormal distributions reveal some common characteristics across levels of J and types of mean configuration. When $f^* = 0$, coverage probability tends to be outside [.925, .975]. When $f^* = .10$, coverage probabilities of the noncentral F distribution-based CI for f^* are all inside the [.94, .96] interval. Coverage probability tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f^* = .25$ with exceptions

occurring principally when data are sampled from the $g = .225$ and $h = .225$ distribution.

Coverage probability of the noncentral F distribution-based CI for f^* tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f^* \geq .40$ and $g = .000$, and $h = .109$. Coverage probability is outside [.925, .975] for only a few cases, all with $J = 6$. Coverage probabilities are mostly outside the [.925, .975] interval when $f^* \geq .40$ for the nonnormal distributions other than the $g = .000$, and $h = .109$ distribution.

Excluding $f^* = 0$, the coverage probability performance of the noncentral F distribution-based CI tends to decline as f^* increases, as the distributions become more long-tailed, and appears to be worse for skewed distributions. Overall, when data are sampled from the $g = 0$ and $h = .109$ distribution more estimates are within the [.925, .975] interval than when data are sampled from the three other nonnormal distributions.

The results of the bootstrap CIs for f^* in Tables 1 to 4 are also fairly similar across levels of J and mean configurations. First, when sampling from the normal distribution, when $f^* = 0$ and $J = 3$, the coverage probabilities of the bootstrap CI for f^* are all above .975. When $f^* = 0$ and $J = 6$, however, they are all inside [.94, .96]. When $f^* = .10$, the coverage probabilities of the bootstrap CI for f^* are all outside [.925, .975] when $J = 3$ and inside [.94, .96] when $J = 6$. When $f^* \geq .25$, coverage probability tends to be inside either [.925, .975] or both intervals.

The coverage probabilities for the bootstrap CI for f^* under non-normality also have some common features across the mean configurations. When $f^* = 0$, the coverage probabilities of the bootstrap CIs for f^* tend to be outside [.925, .975] when $J = 3$ and inside [.94, .96] when $J = 6$. When $f^* = .10$, the coverage probability of the bootstrap CI for f^* tends to be inside [.925, .975] when $J = 6$ except

when $n = 20$ and the mean configuration is equally spaced. Moreover, when $f^* = .10$, the coverage probabilities are mostly inside the [.925, .975] when $J = 3$ with exceptions occurring primarily, but not exclusively, when $g = 0$ and $h = .109$.

Coverage probability tends to be inside either the [.925, .975] interval or both intervals in most conditions when $f^* = .25$ and $J = 3$. When $f^* = .25$ and $J = 6$, more than half of the coverage probabilities are within the [.925, .975] interval. However, under the $g = 0$ and $h = .225$ and $g = .225$ and $h = .225$ data distributions, they are all outside this interval.

Coverage probability of the bootstrap CI for f^* tends to be inside either the [.925, .975] or both intervals in most conditions when $f^* \geq .40$ for the $g = 0$ and $h = .109$ distribution when $J = 3$. However, they have a tendency to be outside the [.925, .975] interval when $J = 6$, especially for the one extreme mean configuration. Coverage probabilities of the bootstrap CI for f^* are mostly outside the [.925, .975] interval when $f^* \geq .40$ for the nonnormal distributions other than $g = .760$ and $h = -.098$. Exceptions occur principally when $f^* = .40$, $J = 3$, and $g = .760$, $h = -.098$ under larger sample sizes ($n = 35$ or 50).

Excluding $f^* = 0$, the coverage probability performance of the bootstrap CI tends to decline as f^* increases, and as the distributions become more long-tailed. As f^* increases, the coverage probability of the bootstrap CI for f^* appears to be worse when $J = 6$ than when $J = 3$. The coverage probability for the bootstrap CI for f^* tends to be poorer than for the noncentral F distribution-based CI both when sampling from normal and nonnormal distributions.

Average Widths of Confidence Intervals for f^*

The average widths of the noncentral F distribution-based and bootstrap CIs for f^* under $J = 3$ and the equally spaced mean

configuration are presented in Table 5. It is observed that generally the average widths of the noncentral F distribution-based CIs are shorter than those of the bootstrap CIs. The difference between the widths of the two CIs becomes smaller as sample size increases. Furthermore, the average width of both type of CIs gets narrower as the sample size increases and the population effect size f^* decreases. Holding f^* and sample size constant, across data distributions, there is very little difference in the width of the noncentral F distribution-based CIs, and there is also very little difference in that of the bootstrap CIs. Presented in Table 6, the average widths of the CIs for f^* under $J = 3$ and the one extreme mean configuration shows little difference from those for the equally spaced mean configuration. This suggests that the type of mean configuration does not substantially affect the width of the CIs and therefore to the precision with which f^* is estimated.

Table 7 shows the average widths of the CIs for f^* under $J = 6$ and the equally spaced mean configuration. It is quite obvious that, when J increases from 3 to 6, the intervals become narrower for all of the combinations of conditions. It is also observed that generally the average widths of the noncentral F distribution-based CIs are shorter than those of the bootstrap CIs. The difference between the widths of the two CIs gets smaller as the sample size increases. In addition, the average widths of both CIs get narrower as the sample size increases and the population ES f^* decreases. Across distributions, there is very little difference in the width of the noncentral F distribution-based CIs and there is also very little difference in that of the bootstrap CIs. The average widths of the CIs for f^* under $J = 6$ and the one extreme mean configuration are presented in Table 8. Again there is little difference between these widths and the widths from those occur for the equally spaced mean configuration, in terms of values as well as patterns observed. This again suggests that the type of mean configuration does not affect the accuracy with which f^* is estimated.

Conclusion

Confidence intervals for the ES have been strongly advocated by statistical methodologists to be used as a useful supplement to and maybe even a superior replacement for the traditional hypothesis testing. However, much investigation is needed to evaluate the robustness of the CIs in order to ensure their proper usage.

In the two group case, it has been reported that in both the independent samples and dependent samples case CIs for Cohen's δ may be misleading because of poor coverage probability when data are nonnormal (Algina & Keselman, 2003b; Algina et al., 2005a, 2006; Kelly, 2005). It has been further reported that the CIs for δ_R , a robust version of δ , have better coverage probability than do CIs for Cohen's δ and that the coverage probability is closer to the nominal level for the percentile bootstrap CIs than for the noncentral t distribution-based CIs (Algina & Keselman, 2003b).

Our study investigated the robustness of the CIs for RMSSE (f^*), in a one-way, fixed-effects, between-subjects ANOVA. The results indicated that the coverage probabilities of the CIs for f^* were not adequate. Under $J = 3$, the probability coverage of the CIs for f^* was acceptable only for (a) CIs constructed by using the noncentral F distribution-based method when data were sampled from the normal distribution and from the $g = .000$ and $h = .109$ distribution, and (b) CIs constructed by using the percentile bootstrap under normality when the population f^* was small ($< .25$). When $J = 6$, the probability coverage of the noncentral F distribution-based CIs was adequate only when the data were normally distributed. The bootstrap CI for f^* provided good probability coverage under normality for almost all values of f^* investigated.

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Table 1.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J = 3, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.976	.984	.974	.981	.978	.984	.974	.977	.978	.980
	35	.965	.984	.973	.987	.972	.984	.976	.985	.978	.985
	50	.976	.990	.978	.990	.979	.985	.975	.985	.971	.983
.10	20	.949	.982	.952	.974	.953	.969	.950	.970	.951	.968
	35	.951	.982	.953	.984	.948	.973	.951	.978	.956	.975
	50	.951	.984	.951	.984	.955	.975	.950	.974	.951	.977
.25	20	.948	.965	.947	.964	.953	.938	.949	.947	.938	.924
	35	.950	.974	.947	.965	.935	.952	.950	.946	.932	.937
	50	.950	.968	.938	.956	.936	.943	.946	.958	.923	.931
.40	20	.942	.948	.941	.946	.925	.917	.933	.918	.913	.876
	35	.959	.956	.932	.938	.926	.923	.935	.925	.912	.894
	50	.951	.950	.932	.935	.908	.912	.934	.926	.900	.891
.55	20	.946	.932	.935	.923	.900	.865	.914	.886	.875	.830
	35	.950	.943	.928	.926	.901	.895	.915	.897	.859	.860
	50	.951	.944	.934	.934	.886	.902	.926	.919	.844	.856
.70	20	.952	.934	.928	.913	.880	.866	.909	.875	.848	.812
	35	.938	.922	.936	.925	.860	.865	.904	.903	.808	.812
	50	.949	.941	.938	.936	.846	.866	.899	.911	.786	.827

Note. Bold values are estimates outside the interval [.94, .96] and bold underlined values are outside the interval [.925, .975].

Table 2.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : $J = 3$, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.973	<u>.984</u>	.970	<u>.978</u>	<u>.980</u>	<u>.982</u>	<u>.986</u>	<u>.983</u>	.974	<u>.980</u>
	35	.975	<u>.991</u>	<u>.977</u>	<u>.989</u>	<u>.978</u>	<u>.986</u>	<u>.977</u>	<u>.986</u>	.974	<u>.982</u>
	50	.972	<u>.986</u>	.972	<u>.983</u>	<u>.976</u>	<u>.986</u>	<u>.976</u>	<u>.986</u>	<u>.977</u>	<u>.986</u>
.10	20	.956	<u>.978</u>	.950	<u>.976</u>	.958	<u>.978</u>	.956	<u>.971</u>	.954	.970
	35	.947	<u>.981</u>	.942	<u>.976</u>	.942	.970	.942	.970	.952	.972
	50	.945	<u>.981</u>	.946	<u>.976</u>	.954	<u>.980</u>	.953	.979	.952	.975
.25	20	.949	.964	.949	.960	.951	.942	.954	.953	.940	.934
	35	.948	.968	.943	.959	.944	.953	.942	.956	.940	.930
	50	.945	.961	.950	.962	.936	.940	.951	.964	.938	.935
.40	20	.948	.954	.938	.936	<u>.920</u>	<u>.899</u>	.933	<u>.922</u>	<u>.911</u>	<u>.886</u>
	35	.950	.950	.942	.939	<u>.922</u>	<u>.919</u>	.933	.929	<u>.912</u>	<u>.899</u>
	50	.950	.950	.942	.944	<u>.916</u>	<u>.918</u>	.934	.933	<u>.896</u>	<u>.894</u>
.55	20	.945	.936	.933	.927	<u>.908</u>	<u>.876</u>	.931	<u>.907</u>	<u>.881</u>	<u>.850</u>
	35	.944	.938	.928	<u>.922</u>	<u>.892</u>	<u>.880</u>	<u>.916</u>	<u>.912</u>	<u>.867</u>	<u>.864</u>
	50	.949	.945	.935	.930	<u>.885</u>	<u>.889</u>	<u>.923</u>	.928	<u>.836</u>	<u>.862</u>
.70	20	.949	.932	.940	<u>.921</u>	<u>.871</u>	<u>.845</u>	<u>.910</u>	<u>.888</u>	<u>.843</u>	<u>.811</u>
	35	.945	.935	.934	.937	<u>.850</u>	<u>.851</u>	<u>.896</u>	<u>.894</u>	<u>.807</u>	<u>.822</u>
	50	.950	.941	.936	.936	<u>.856</u>	<u>.867</u>	<u>.905</u>	<u>.922</u>	<u>.791</u>	<u>.828</u>

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Table 3.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* : J = 6, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	<u>.977</u>	.950	.975	.944	<u>.980</u>	.949	<u>.976</u>	.946	<u>.980</u>	.948
	35	.974	.955	<u>.976</u>	.950	<u>.976</u>	.953	.975	.949	<u>.980</u>	.954
	50	.972	.956	.972	.954	<u>.985</u>	.966	<u>.977</u>	.952	<u>.978</u>	.956
.10	20	.953	.951	.952	.927	.954	.935	.959	.932	.952	<u>.924</u>
	35	.951	.943	.942	.944	.953	.938	.950	.943	.944	.928
	50	.944	.945	.950	.945	.958	.946	.940	.942	.954	.928
.25	20	.948	.927	.948	<u>.921</u>	.938	<u>.892</u>	.950	<u>.905</u>	.928	<u>.871</u>
	35	.952	.944	.944	.933	.937	<u>.905</u>	.938	<u>.910</u>	<u>.919</u>	<u>.889</u>
	50	.954	.954	.943	.933	.932	<u>.910</u>	.944	<u>.926</u>	<u>.910</u>	<u>.880</u>
.40	20	.950	.933	.945	<u>.920</u>	<u>.917</u>	<u>.858</u>	<u>.922</u>	<u>.901</u>	<u>.880</u>	<u>.819</u>
	35	.953	.937	.940	.927	<u>.900</u>	<u>.877</u>	.928	<u>.906</u>	<u>.860</u>	<u>.837</u>
	50	.955	.947	.943	.935	<u>.904</u>	<u>.890</u>	.932	<u>.924</u>	<u>.860</u>	<u>.859</u>
.55	20	.949	<u>.923</u>	.934	<u>.904</u>	<u>.876</u>	<u>.825</u>	<u>.914</u>	<u>.874</u>	<u>.856</u>	<u>.800</u>
	35	.958	.940	.931	<u>.921</u>	<u>.872</u>	<u>.860</u>	<u>.902</u>	<u>.889</u>	<u>.818</u>	<u>.807</u>
	50	.954	.939	.930	.928	<u>.869</u>	<u>.884</u>	<u>.914</u>	<u>.910</u>	<u>.808</u>	<u>.840</u>
.70	20	.955	.930	.932	<u>.893</u>	<u>.849</u>	<u>.816</u>	<u>.893</u>	<u>.876</u>	<u>.790</u>	<u>.752</u>
	35	.942	.930	<u>.923</u>	<u>.914</u>	<u>.826</u>	<u>.837</u>	<u>.892</u>	<u>.893</u>	<u>.766</u>	<u>.784</u>
	50	.943	.932	<u>.918</u>	.927	<u>.820</u>	<u>.857</u>	<u>.895</u>	<u>.918</u>	<u>.752</u>	<u>.823</u>

Note. Bold values are estimates outside the interval [.94, .96] and bold underlined values are outside the interval [.925, .975].

Table 4.
Estimated coverage probabilities for nominal 95% noncentral F distribution-based and percentile
bootstrap (boot) CIs for f^* : J = 6, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.975	.949	<u>.976</u>	.946	.974	.942	.972	.936	<u>.982</u>	.948
	35	<u>.976</u>	.956	.968	.947	<u>.978</u>	.951	<u>.978</u>	.952	<u>.976</u>	.954
	50	.975	.958	<u>.976</u>	.959	<u>.982</u>	.958	<u>.980</u>	.956	.970	.944
.10	20	.949	.944	.948	.927	.959	.938	.954	.926	.949	.926
	35	.954	.950	.954	.943	.948	.932	.947	.930	.955	.935
	50	.952	.948	.946	.950	.950	.936	.957	.945	.953	.933
.25	20	.954	.934	.947	<u>.920</u>	.935	<u>.894</u>	.941	<u>.911</u>	.942	<u>.884</u>
	35	.953	.940	.948	.946	.939	<u>.910</u>	.947	.933	.927	<u>.888</u>
	50	.953	.947	.945	.929	.932	<u>.909</u>	.948	.939	<u>.917</u>	<u>.898</u>
.40	20	.952	.930	.951	<u>.924</u>	<u>.918</u>	<u>.860</u>	.947	<u>.898</u>	<u>.890</u>	<u>.827</u>
	35	.946	.932	.937	<u>.924</u>	<u>.911</u>	<u>.892</u>	.934	<u>.917</u>	<u>.883</u>	<u>.862</u>
	50	.950	.936	.938	.931	<u>.900</u>	<u>.894</u>	.932	.932	<u>.856</u>	<u>.860</u>
.55	20	.955	.931	.938	<u>.902</u>	<u>.877</u>	<u>.838</u>	<u>.923</u>	<u>.886</u>	<u>.844</u>	<u>.793</u>
	35	.951	.930	.929	<u>.919</u>	<u>.863</u>	<u>.862</u>	<u>.916</u>	<u>.909</u>	<u>.821</u>	<u>.824</u>
	50	.949	.936	<u>.922</u>	.925	<u>.858</u>	<u>.879</u>	<u>.909</u>	<u>.908</u>	<u>.783</u>	<u>.820</u>
.70	20	.945	<u>.915</u>	.929	<u>.893</u>	<u>.848</u>	<u>.826</u>	<u>.914</u>	<u>.885</u>	<u>.794</u>	<u>.754</u>
	35	.947	.935	<u>.920</u>	<u>.911</u>	<u>.828</u>	<u>.834</u>	<u>.896</u>	<u>.908</u>	<u>.752</u>	<u>.790</u>
	50	.942	.930	.926	<u>.918</u>	<u>.817</u>	<u>.849</u>	<u>.902</u>	<u>.920</u>	<u>.740</u>	<u>.815</u>

Note. Bold values are estimates outside the interval $[.94, .96]$ and bold underlined values are outside the interval $[.925, .975]$.

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Table 5.

Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=3, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.446	.534	.449	.529	.442	.515	.451	.520	.448	.513
	35	.338	.393	.334	.388	.339	.383	.335	.381	.340	.381
	50	.281	.323	.278	.321	.283	.318	.284	.320	.285	.316
.10	20	.467	.551	.470	.545	.479	.542	.476	.541	.479	.542
	35	.367	.416	.361	.409	.367	.408	.369	.409	.370	.407
	50	.309	.346	.313	.348	.314	.345	.314	.345	.317	.346
.25	20	.560	.628	.561	.627	.568	.627	.568	.629	.577	.640
	35	.453	.495	.452	.490	.457	.493	.456	.492	.457	.497
	50	.395	.425	.393	.424	.396	.427	.396	.426	.394	.429
.40	20	.641	.701	.641	.702	.642	.710	.638	.707	.648	.724
	35	.497	.533	.495	.533	.495	.547	.497	.543	.496	.555
	50	.413	.437	.413	.442	.413	.454	.414	.449	.413	.465
.55	20	.676	.726	.676	.739	.678	.764	.676	.754	.677	.781
	35	.504	.526	.504	.538	.506	.569	.505	.559	.507	.593
	50	.417	.429	.418	.444	.419	.477	.418	.461	.420	.498
.70	20	.693	.732	.692	.753	.696	.813	.696	.800	.702	.842
	35	.514	.527	.515	.550	.517	.612	.516	.590	.521	.640
	50	.428	.433	.428	.457	.430	.512	.428	.493	.433	.547

Table 6.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=3, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.448	.535	.452	.531	.453	.519	.442	.514	.452	.518
	35	.335	.392	.336	.387	.338	.383	.334	.381	.341	.380
	50	.280	.324	.287	.325	.283	.316	.280	.317	.284	.316
.10	20	.472	.552	.473	.549	.482	.546	.473	.539	.476	.540
	35	.361	.413	.365	.410	.369	.410	.367	.408	.371	.410
	50	.312	.349	.312	.346	.315	.346	.315	.345	.319	.348
.25	20	.562	.629	.566	.629	.573	.634	.564	.622	.578	.637
	35	.452	.493	.456	.493	.457	.498	.455	.489	.464	.503
	50	.394	.423	.394	.423	.395	.426	.395	.422	.400	.432
.40	20	.641	.703	.638	.698	.643	.713	.643	.701	.645	.716
	35	.496	.534	.496	.533	.496	.541	.496	.532	.496	.549
	50	.414	.437	.414	.440	.414	.456	.414	.442	.413	.459
.55	20	.676	.726	.675	.737	.678	.763	.679	.744	.679	.777
	35	.504	.527	.504	.542	.506	.568	.505	.551	.507	.586
	50	.417	.428	.418	.443	.420	.477	.418	.453	.420	.493
.70	20	.692	.729	.693	.756	.698	.805	.695	.782	.701	.840
	35	.514	.530	.514	.553	.518	.612	.516	.581	.521	.634
	50	.427	.433	.427	.457	.430	.515	.428	.484	.432	.544

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Table 7.

Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=6, equally spaced mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.321	.434	.325	.433	.319	.427	.318	.428	.324	.428
	35	.239	.320	.242	.319	.241	.317	.241	.317	.241	.315
	50	.202	.266	.201	.265	.202	.263	.200	.263	.201	.263
.10	20	.344	.446	.350	.446	.348	.443	.350	.443	.351	.443
	35	.274	.337	.271	.335	.277	.336	.272	.334	.276	.333
	50	.238	.284	.237	.283	.239	.283	.238	.282	.238	.280
.25	20	.426	.479	.424	.475	.428	.473	.429	.474	.429	.474
	35	.332	.353	.332	.351	.331	.349	.332	.350	.331	.350
	50	.275	.285	.275	.286	.273	.285	.275	.286	.273	.287
.40	20	.450	.469	.448	.468	.446	.475	.447	.473	.444	.481
	35	.323	.330	.324	.334	.324	.347	.324	.341	.324	.357
	50	.265	.268	.265	.274	.265	.290	.265	.281	.266	.301
.55	20	.442	.452	.442	.463	.444	.492	.443	.480	.444	.510
	35	.324	.328	.324	.339	.325	.374	.325	.356	.326	.393
	50	.268	.270	.268	.282	.269	.317	.269	.298	.270	.340
.70	20	.448	.457	.449	.478	.453	.529	.449	.513	.455	.565
	35	.332	.336	.332	.357	.334	.412	.333	.386	.336	.438
	50	.276	.277	.276	.299	.277	.352	.276	.323	.278	.381

Table 8.
Average widths of noncentral F distribution-based (NCF) and percentile bootstrap (boot) CIs for f^* :
J=6, one extreme mean configuration

f^*	n	Normal		$g = .000$ $h = .109$		$g = .000$ $h = .225$		$g = .760$ $h = -.098$		$g = .225$ $h = .225$	
		NCF	boot	NCF	boot	NCF	boot	NCF	boot	NCF	boot
.00	20	.322	.434	.319	.431	.322	.429	.323	.432	.321	.427
	35	.239	.319	.239	.318	.243	.317	.242	.318	.241	.315
	50	.202	.266	.199	.263	.202	.264	.199	.263	.203	.263
.10	20	.350	.449	.347	.444	.349	.442	.352	.444	.355	.445
	35	.274	.338	.274	.337	.275	.333	.274	.334	.278	.335
	50	.236	.284	.237	.283	.236	.281	.237	.282	.240	.282
.25	20	.425	.480	.425	.477	.427	.474	.429	.469	.432	.473
	35	.333	.354	.331	.351	.332	.351	.333	.344	.331	.345
	50	.276	.285	.275	.286	.274	.286	.276	.278	.274	.284
.40	20	.449	.469	.449	.468	.447	.475	.449	.453	.446	.475
	35	.324	.329	.324	.335	.324	.350	.323	.330	.324	.352
	50	.265	.268	.265	.274	.265	.290	.265	.273	.266	.296
.55	20	.442	.452	.442	.463	.444	.496	.443	.468	.445	.508
	35	.324	.327	.324	.340	.325	.375	.324	.349	.326	.391
	50	.268	.270	.268	.283	.269	.318	.269	.294	.270	.334
.70	20	.448	.458	.449	.481	.452	.534	.449	.502	.455	.557
	35	.332	.334	.332	.357	.335	.412	.333	.380	.336	.438
	50	.276	.277	.276	.297	.278	.350	.276	.320	.278	.381

However, for all other combinations of conditions, the bootstrap CI did not provide accurate probability coverage. Furthermore, excluding $f^* = 0$, the coverage performance of the noncentral F distribution-based CIs tended to decline as f^* increased, as the distributions became more long-tailed, and appeared to be worse for skewed distributions. Overall, the noncentral F distribution-based CIs for f^* yielded relatively better probability coverage than that of the bootstrap CIs for f^* . The type of mean configurations and the number of treatment groups did not appear to affect the coverage probability of the CIs for f^* considerably. Therefore, the coverage performance of the CIs for f^* might be generalizable over types of mean configuration and various numbers of treatment groups.

The widths of the noncentral F distribution-based CIs for f^* were all narrower than those of the bootstrap CIs under the same condition. The interval widths of the CIs for f^* were relatively unchanged across data distributions. The width of both estimated CIs became narrower as the number of levels for J increased, the sample size increased, and the population effect size f^* decreased.

In summary, both the noncentral F distribution-based and the bootstrap CIs for f^* , which are based on the least-square estimators, yielded inadequate coverage probabilities. Thus an important task to help researchers who want to set a CI around f^* is developing a better interval than the noncentral F distribution-based or percentile bootstrap CI. An improved measure of effect size might be attained by substituting robust estimators, e.g., trimmed means and Winsorized variances, for the least-square values. Thus, one of our future studies has set out to propose a robust version of f^* . A robust measure of effect size may yield better coverage probabilities and provide a measure that is not likely to be strongly affected by outlying data points.

References

- Algina, J., & Keselman, H. J. (2003a). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement*, 63, 537-553.
- Algina, J., & Keselman, H. J. (2003b, May). Confidence Intervals for Effect Sizes. Paper presented at a conference in honor of H. Swaminathan, Amherst, MA.
- Algina, J., Keselman, H. J., & Penfield, R. D. (2005a). Effect sizes and their intervals: the two-level repeated measures case. *Educational and Psychological Measurement*, 65, 241-258.
- Algina, J., Keselman, H. J., & Penfield, R. D. (2005b). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317-328.
- Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size Statistic. *Educational and Psychological Measurement*, 66, 945-990.
- American Psychological Association. (2001). *Publication manual of the American Psychological Association* (5th ed.). Washington, DC.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, 31, 144-152.
- Cliff, N. (1993). Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychological Bulletin*, 114, 494-509.
- Cliff, N. (1996). Answering ordinal questions with ordinal data using ordinal statistics, *Multivariate Behavioral Research*, 31, 331-350.
- Cohen, J. (1965). Some statistical issues in psychological research. In B.B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95-121). New York: Academic Press.
- Cohen, J. (1969). *Statistical power analyses for the behavioral sciences*. NY: Academic Press, Inc.
- Cohen, J. (1988). *Statistical power and analysis for the behavioral sciences* (2nd ed.). NY: Academic Press.

- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologists*, 49, 997-1003.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61, 532-574.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60, 178-180.
- Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. NY: Chapman and Hall.
- Finch, S., Thomason, N., & Cumming, G. (2002). Past and future American Psychological Association guidelines for statistical practice. *Theory and Psychology*, 36, 312-324.
- Grissom, R. J., & Kim, J. J. (2005). Effect sizes for research: A broad practical approach. Mahwah, NJ: Lawrence Erlbaum.
- Hays, W. L. (1963). *Statistics*. NY, Rinehart & Winston.
- Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.
- Hess, M. R. & Kromrey, J. D. (April, 2003). *Confidence intervals for standardized mean differences: An empirical comparison of bootstrap methods under non-normality and heterogeneous variances*. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.
- Hess, M. R. & Kromrey, J. D. (April, 2004). *Robust confidence intervals for effect sizes: A comparative study of Cohen's d and Cliff's delta under non-normality and heterogeneous variances*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.
- Hoaglin, D. C. (1983). Summarizing shape numerically: The g-and h distributions. In D. D. Hoaglin, F. Mosteller, & Tukey, J. W. (Eds.), *Data analysis for tables, trends, and shapes: Robust and exploratory techniques*. New York: Wiley.
- Hogarty, K. Y., & Kromrey, J. D. (April, 2001). *We've been reporting some effect sizes: Can you guess what they mean?* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.
- Johnson, N. L., & Welch, B. L. (1940). Applications of the non-central t distribution. *Biometrika*, 34, 362-389.
- Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65, 51-69.
- Kraemer, H. C., & Andrews, G. A. (1982). A nonparametric technique for meta-analysis effect size calculation. *Psychological Bulletin*, 91, 404-412.
- Martinez, J., & Iglewicz, B. (1984). Some properties of the Tukey g and h family of distributions. *Communications in Statistics, Theory and Methods*, 13, 353-369.
- McGraw, K. O. & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, 111, 361-365.
- Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34, 103-115.
- Murphy, K. R. (1997). Editorial. *Journal of Applied Psychology*, 82, 3-5.
- Nickerson, R. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8, 434-447.
- SAS Institute Inc. (1999). *SAS/IML user's guide, version 8*. Cary, NC: Author.
- Serlin, R. C., & Lapsley, D. K. (1985). Rationality in psychological research: The good enough principle. *American Psychologist*, 40, 73-83.
- Steiger, J. H. (2004). Beyond the F test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164-182.

STANDARDIZED EFFECT SIZE IN ONE-WAY FIXED-EFFECTS ANOVA

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. Harlow, S. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Hillsdale, NJ: Erlbaum.

Thompson, B. (1994). Guidelines for authors. *Educational and Psychological Measurement*, 54, 837-847.

Vargha, A. & Delaney, H. D. (2000). A critique and improvement of the CL Common Language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, 25, 101-132.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego: Academic Press.

Wilcox, R. R. & Keselman. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods*, 8, 254-274.

Wilcox, R. R. & Muska, J. (1999). Measuring effect size: A non-parametric analogue of ω^2 . *British Journal of Mathematical and Statistical Psychology*, 52, 93-110.

Wilkinson, L. and the Task force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist*, 54, 594-604.

An Evaluation of Standard, Alternative, and Robust Slope Test Strategies

Tim Moses
Educational Testing Service

Alan Klockars
University of Washington

The robustness and power of nine strategies for testing the differences between two groups' regression slopes under nonnormality and residual variance heterogeneity are compared. The results showed that three most robust slope test strategies were the combination of the trimmed and Winsorized slopes with the James second order test, the combination of Theil-Sen with James, and Theil-Sen with percentile bootstrapping. The slope tests based on Theil-Sen slopes were more powerful than those based on trimmed and Winsorized slopes.

Key words: slopes, least squares, Theil-Sen, robust regression, James second order, nonnormality, residual variance heterogeneity

Introduction

The question of whether group differences are constant or vary across levels of an individual difference variable (X) has been considered in many fields of social science, including clinical psychology (Dance & Neufeld, 1988), organizational research (Aguinis & Pierce, 1998; Hunter, Schmidt, & Hunter, 1979), learned helplessness (Seligman, 2002) and education's search for Aptitude-Treatment Interactions (Cronbach & Snow, 1977). Several strategies have been proposed for evaluating the consistency of group differences across X based on fitting regression lines that predict outcome Y from X in separate treatment groups and then conducting a significance test for the homogeneity of the groups' regression slopes.

Tim Moses is a Senior Psychometrician at Educational Testing Service where he works on several testing programs. Tim completed his PhD in Educational Psychology at the University of Washington. E-mail him at mosest@optonline.net. Alan Klockars is Professor of Educational Psychology at the University of Washington. His research concerns multiple comparisons and, more recently, methods of conducting ATI research. E-mail him at klockars@u.washington.edu.

The purpose of this study is to compare some of the recently-researched methods of slope estimation and testing under conditions of nonnormality and residual variance heterogeneity.

The slope test strategies considered in this study are approaches to estimating the following model,

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + \varepsilon_{ij}, \quad (1)$$

where outcome Y for individual i (= 1 to N) in group j (= 1 to J) is a linear function of a continuous X, β_{0j} and β_{1j} are the population intercepts and slopes of the regression line for each of J groups, and the ε_{ij} are the residuals. The strategies for assessing differences in the β_{1j} 's reviewed below are most easily understood in terms of alternative expressions of (1). When J = 2, (1) can be expressed as,

$$Y_{ij} = \beta_0 + \beta_1X_{ij} + \beta_2G_{ij} + \beta_3X_{ij}G_{ij} + \varepsilon_{ij}, \quad (2)$$

where G_{ij} is a dichotomously-coded group membership variable. A more general matrix version of (1) and (2) is,

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \quad (3)$$

where $\mathbf{Y}$ is an N by 1 column vector, $\mathbf{X}$ is an N by K “design matrix” corresponding to the K β ’s (including a column of 1’s for estimating β_0), $\boldsymbol{\beta}$ is a K by 1 column vector of β ’s and $\mathbf{e}$ is an N by 1 column vector of residuals.

Standard slope estimation and slope test

The standard slope test uses “least squares” estimates of the β ’s (i.e., $\hat{\beta}$ ’s) that minimize the sum of the squared residuals, $\hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Because $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ is a convex function of $\hat{\boldsymbol{\beta}}$, it can be minimized by differentiating with respect to $\hat{\boldsymbol{\beta}}$, setting this derivative to zero and solving for $\hat{\boldsymbol{\beta}}$, resulting in the closed form solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}. \quad (4)$$

Equivalently, group j ’s slope can be estimated

$$\hat{\beta}_{1j} = \frac{\sum_{i \in j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sum_{i \in j} (x_{ij} - \bar{x}_j)^2}, \quad (5)$$

where $\bar{x}_j$ and $\bar{y}_j$ are the means of X and Y in group j .

The standard test for assessing the differences of J slopes is an F test,

$$F_{\text{SlopesStandard}} = \frac{\left(\frac{1}{J-1} \right) \sum_j \left(\hat{\beta}_{1j}^2 - \hat{\beta}_{1,\text{Standard}}^2 \right) \sum_{i \in j} (x_{ij} - \bar{x}_j)^2}{\left(\frac{1}{N-2J} \right) \sum_j (N_j - 2) \hat{\sigma}_{ej}^2} \quad (6),$$

where $\hat{\sigma}_{ej}^2 = \frac{\sum_{i \in j} (\hat{e}_{ij})^2}{N_j - 2}$ and

$$\hat{\beta}_{1,\text{Standard}} = \frac{\sum_j \hat{\beta}_{1j} \sum_{i \in j} (x_{ij} - \bar{x}_j)^2}{\sum_j \sum_{i \in j} (x_{ij} - \bar{x}_j)^2} \quad \text{is the variance-}$$

weighted common slope (Myers & Well, 1995, p. 421-422). (6) is evaluated on an F distribution with $J-1$ and $N-2J$ degrees of freedom. With $J=2$, a t -test of $\hat{\beta}_3$ in model (2) that is equivalent to the F test in (6) can be conducted by obtaining

the standard error of $\hat{\beta}_3$ as the square root of one of the diagonal elements in the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, $\frac{\hat{\mathbf{e}}'\hat{\mathbf{e}}}{N-K}(\mathbf{X}'\mathbf{X})^{-1}$, and

evaluating $\frac{\hat{\beta}_3}{\text{SE}(\hat{\beta}_3)}$ on a t distribution with

$N-K = N-4$ degrees of freedom. The referencing of the standard test statistics to F and t distributions is justified when the data meet particular assumptions, namely that the ε_{ij} are normally and independently distributed with equal variances across the J groups.

The standard methods for estimating and testing slopes are problematic when data are nonnormal and residual variances are heterogeneous (Conover & Iman, 1982; Conerly & Mansfield, 1988; Headrick & Sawilowsky, 2000; Klockars & Moses, 2002; Dretzke, Levin & Serlin, 1982; Overton, 2001; Alexander & Deshon, 1994; Deshon & Alexander, 1996). When distributions exhibit heavy-tailed nonnormality, extreme scores occur more often than when distributions are normal, increasing the variability of the estimated slopes, reducing the estimated standard errors, and making the standard test excessively liberal. When groups’ residual variances and sample sizes differ, the standard test’s pooling of groups’ residual variances, $\left(\frac{1}{N-2J} \right) \sum_j (N_j - 2) \hat{\sigma}_{ej}^2$, is

problematic, making the standard slope test either liberal or conservative depending on whether the larger and smaller group has the larger or smaller residual variance. The inaccuracy of the standard test is disturbing given that nonnormality and residual variance heterogeneity appear to be common in actual data (Micceri, 1989; Aguinis, Peterson & Pierce, 1999). What follows are detailed definitions of slope test strategies that may outperform the standard test when distributions are nonnormal and residual variances are heterogeneous.

Slope tests for nonnormal data: Central tendency strategies

Two approaches to slope estimation view group j ’s slope in (5),

$$\hat{\beta}_{1j} = \frac{\sum_{i \in j} (x_{ij} - \bar{x}_j)(y_{ij} - \bar{y}_j)}{\sum_{i \in j} (x_{ij} - \bar{x}_j)^2}, \text{ as a central value of}$$

the slopes that can be created from pairs of observations in the data,

$$b_{1,ij} = \frac{(y_{ij} - y_{i'j})}{(x_{ij} - x_{i'j})}, i \neq i', x_{ij} \neq x_{i'j}, \text{ and then try}$$

reduce the influence of the extreme observations on the central value. These ‘central tendency’ approaches define extreme observations in terms of both X and Y, so that the screening of extreme observations caused by nonnormality could potentially address slope estimation problems such as leverage (observations that are extreme on X), discrepancy (observations that are extreme with respect to the regression line), and outliers on Y. One popular strategy is the Theil-Sen slope estimator (Theil, 1950; Sen, 1968; Wilcox, 2004; Wilcox & Keselman, 2004; Ebrahem & Al-Nasser, 2005; Wang, 2005). The Theil-Sen estimate is the median of the slopes that can be computed from the $N_j(N_j-1)/2$ pairs of observations in the data. Percentile bootstrapping methods can be used to test for differences between groups’ Theil-Sen slopes (i.e., draw 599 random samples with replacement from the $J = 2$ datasets, compute the differences in Theil-Sen slopes in each of these datasets, and determine if the middle $(1-\alpha)\%$ of the 599 slope differences contain zero, Wilcox, 2005).

A less-familiar alternative to the Theil-Sen slope estimate is the application of the trimming and Winsorizing strategies that are typically proposed in tests of mean differences

to $\frac{(y_{ij} - y_{i'j})}{(x_{ij} - x_{i'j})}$ (Guo, 1996; Luh & Guo, 2000). To

obtain trimmed and Winsorized estimates of slopes and their variances, rank order the x ’s in each of the J groups, $x_{1j} < x_{2j} < \dots < x_{N_jj}$. When the number of observations in group j is even ($N_j=2m_j$) consider m_j independent slope estimates,

$$b_{1,i+m_j,j} = \frac{(y_{i+m_j,j} - y_{ij})}{(x_{i+m_j,j} - x_{ij})}. \quad (7)$$

When the number of observations in j is odd ($N_j=2m_j+1$), a pooling is done so that

observations $y_{2m_j,j}$ and $y_{2m_j+1,j}$ are pooled, $x_{2m_j,j}$ and $x_{2m_j+1,j}$ are pooled, and $y_{2m_j,j}$ and $x_{2m_j,j}$ are replaced by $y_{2m_j,j} = (y_{2m_j,j} + y_{2m_j+1,j})/2$ and $x_{2m_j,j} = (x_{2m_j,j} + x_{2m_j+1,j})/2$.

The trimming and Winsorizing is done for each of the j slopes and standard errors. Let $g_j = \gamma m_j$ where γ represents the proportion of observations to be trimmed from each tail of the ordered distribution of

$b_{1,l,j}, l = 1 \text{ to } m_j, b_{1,l,j} \leq b_{1,2,j} \leq \dots \leq b_{1,m_j,j}$. Let $h_j = m_j - 2g_j$ be the effective sample size after trimming.

The trimmed mean slope is computed as

$$\bar{b}_{1,j} = \frac{\sum_{l=g_j+1}^{m_j-g_j} b_{1,l,j}}{h_j}. \quad (8)$$

Winsorized slope observations are obtained by,

$$bw_{1,l,j} = \begin{cases} b_{1,(g_j+1)j} & \text{if } b_{1,l,j} \leq b_{1,(g_j+1)j} \\ b_{1,l,j} & \text{if } b_{1,(g_j+1)j} < b_{1,l,j} < b_{1,(m_j-g_j)j} \\ b_{1,(m_j-g_j)j} & \text{if } b_{1,l,j} \geq b_{1,(m_j-g_j)j} \end{cases}$$

(9). The variance of the trimmed mean slope is computed as a function of the Winsorized variance,

$$\sigma_{bwj}^2 = \frac{1}{h_j(h_j-1)} \sum_{l=1}^{m_j} (bw_{1,l,j} - \frac{\sum bw_{1,l,j}}{m_j})^2. \quad (10)$$

To assess the differences in trimmed slopes, replace the $\hat{\beta}_{1j}$ in (6) with $\bar{b}_{1,j}$, the

$$\left(\frac{1}{N-2J} \right) \sum_j (N_j-2) \hat{\sigma}_{ej}^2 \quad \text{with}$$

$$\left(\frac{1}{\sum_j h_j - J} \right) \sum_j h_j (h_j - 1) \sigma_{bwj}^2, \quad \text{and the}$$

$\sum_{i \in j} (X_{ij} - \bar{X}_j)^2$ with h_j . These replacements to

(6) cause the standard test of slope differences to resolve into an F test for independent trimmed means with $J-1$ and $\sum_j h_j - J$ degrees of freedom,

$$F_{\text{SlopesTrimmed}} = \frac{\left(\frac{1}{J-1} \sum_j h_j \left(\bar{b}_{ij}^2 - \frac{\left(\sum_j h_j \bar{b}_{ij} \right)^2}{\sum_j h_j} \right) \right)}{\left(\frac{1}{\sum_j h_j - J} \sum_j h_j (h_j - 1) \sigma_{b_{ij}}^2 \right)}. \quad (11)$$

Slope tests for nonnormal data: MM Regression

In “minimum maximum likelihood type” (MM, Yohai, 1987) regression, extreme observations are addressed in the minimization process used to estimate the regression line. While the standard slope estimation process is based on minimizing the sum of all squared residuals, the robust regression paradigm views the least squares approach as one of several possible functions, ξ , of the scaled residuals that could be minimized,

$$\sum_j \sum_{i \text{ in } j} \xi \left(\frac{\varepsilon_{ij}}{\sigma} \right). \quad (12)$$

Some choices of ξ can produce β estimates that outperform the standard method's β 's in terms of their “breakdown” rates (i.e., the smallest percentage of contaminated observations needed to render $\hat{\beta}$ useless). One popular ξ (SAS, 2003) is the Tukey weight function,

$$\xi(s) = \begin{cases} 3\left(\frac{s}{\kappa}\right)^2 - 3\left(\frac{s}{\kappa}\right)^4 + \left(\frac{s}{\kappa}\right)^6 & \text{if } |s| \leq \kappa, \\ 1 & \text{otherwise.} \end{cases} \quad (13)$$

In (13), κ is a constant selected to obtain desirable properties. A κ value of 3.44 results in parameter estimates that are 85% as efficient as least squares estimates when the data are normal (Holland & Welsh, 1977). When data contain outliers that are discrepant with respect to the regression line, κ defines a range around which the observations outside of the range have reduced contribution to the slope estimates.

The search for β 's that minimize (12) is similar to the standard test in that β_k 's are found such that the derivatives of (12) with respect to

the β_k 's are zero,

$$\sum_j \sum_{i \text{ in } j} \frac{\partial \xi}{\partial s} (s_{ijk}) x_{ijk} = 0, k = 1 \text{ to } K. \quad \text{Unlike the}$$

least squares estimation methods used with the standard test, with MM regression there are no closed-form solutions to minimizing (12). The following is an outline of the three-stage MM algorithm for estimating the β_k 's.

The first step of MM regression is to obtain robust starting values for the β_k 's and σ . The current SAS procedure for MM uses Least Trimmed Squares estimates as starting values (Rousseeuw, 1984; SAS Institute, 2003). The basic idea of LTS estimation is to draw samples of K observations from the N total observations in the data set. In each sample, obtain least squares estimates of the β_k 's and find the ones

that minimize $\sum_i^h (\varepsilon_i)^2$, where $h = \frac{3N + K + 1}{4}$

and observations i through h reference the h smallest squared residuals. Additional features of the LTS algorithm involve intercept adjustments that reduce $\sum_i^h (\varepsilon_i)^2$ and

computational search processes designed to find final β_k estimates quickly in extremely large datasets (Rousseeuw & Van Driessen, 2000). One preliminary estimate of σ is computed as,

$$s_{\text{LTS}} = d \sqrt{\frac{1}{h} \sum_i^h (\varepsilon_i)^2}, \quad (14)$$

where $d = 1/\sqrt{1 - \frac{2N}{hc} \phi(1/c)}$, $c = 1/\Phi(\frac{h+N}{2N})$, and Φ and ϕ are the cumulative and probability density functions of the standard normal distribution.

A more efficient estimate of σ than s_{LTS} can also be computed,

$$W_{\text{scale}} = \sqrt{\frac{\sum_i^N w_i (\varepsilon_i)^2}{\sum_i w_i - K}}, \quad (15)$$

where $w_i = \begin{cases} 0 & \text{if } |\varepsilon_i|/s_{\text{LTS}} > 3 \\ 1 & \text{otherwise} \end{cases}$.

With initial estimates of the β_k 's and σ , the second step is to conduct iterative calculations to produce a converged σ value,

$$(\sigma^{m+1})^2 = \frac{1}{(N-K) \int \xi(s) \partial \Phi(s)} \sum_i^N \xi\left(\frac{\varepsilon_i}{\sigma^m}\right) (\sigma^m)^2, \quad (16)$$

where $\int \xi(s) \partial \Phi(s)$ denotes an expected value of $\xi(s)$ when the s are from a normal distribution (about .25 for the Tukey bisquare $\xi(s)$ with $\kappa = 2.9366$). In (16), setting $\kappa = 2.9366$ results in the σ having a breakdown rate of 25% (SAS, 2003).

The third step is to conduct an iterative search for a final solution of the β_k 's with a fixed σ value

$$\beta^{m+1} = (\mathbf{X}^t \mathbf{\Omega} \mathbf{X})^{-1} \mathbf{X}^t \mathbf{\Omega} \mathbf{Y}, \quad (17)$$

where $\mathbf{\Omega}$ is an N by N matrix with diagonal entries $\frac{\partial \xi(s)}{\partial s} \frac{1}{s}$ where the s are the scaled residuals from the m th iteration step and $\kappa = 3.44$ in default SAS routines (SAS, 2003). The entries for $\mathbf{\Omega}$ are the “reweighted” part of MM’s iteratively reweighted least squares algorithm, and for the Tukey $\xi(s)$ given in (13) are known as the Tukey bisquare weight function.

At convergence, there are several estimates of the asymptotic variance-covariance matrix of β (SAS, 2003). One version is,

$$\left(1 + \frac{K}{N \left(\frac{1}{N} \sum_i \partial^2 \xi(\varepsilon_i) / \partial^2 \varepsilon \right)^2} \right) \frac{\sigma^2 (\partial^2 \xi(\varepsilon) / \partial^2 \varepsilon)}{\left(\frac{1}{N} \sum_i \partial^2 \xi(\varepsilon_i) / \partial^2 \varepsilon \right)^2} \frac{(1/(N-K) \sum_i \partial \xi(\varepsilon_i) / \partial \varepsilon)^2}{\left(\frac{1}{N} \sum_i \partial^2 \xi(\varepsilon_i) / \partial^2 \varepsilon \right)} \mathbf{W}^1, \quad (18)$$

where $\left(1 + \frac{K}{N \left(\frac{1}{N} \sum_i \partial^2 \xi(\varepsilon_i) / \partial^2 \varepsilon \right)^2} \right)$ is a

correction factor, $\partial^2 \xi(\varepsilon) / \partial^2 \varepsilon$ is the second derivative of ξ with respect to the residuals, and

$\mathbf{W}$ is a K by K matrix with entries $\mathbf{W}_{kk'} = \sum_i (\partial^2 \xi(\varepsilon_i) / \partial^2 \varepsilon_i) x_{ik} x_{ik'}$.

The preceding review provides some insight into the kinds of nonnormality problems for which MM might be especially useful, which are probably situations with outliers that do not “mask” themselves by exerting heavy influence on the regression line. Many of the steps of the MM estimation process are analogues to the standard method’s estimation, including the use of least squares estimation used in the LTS starting values, the computation of the β_k 's (equation 17 is a weighted version of equation 4), and the computation of the MM standard errors ($\mathbf{W}$ in equation 18 is a weighted version

of $(\mathbf{X}^t \mathbf{X})$ in $\frac{\hat{\mathbf{e}}^t \hat{\mathbf{e}}}{N-K} (\mathbf{X}^t \mathbf{X})^{-1}$). The relatedness of MM computations to the standard method’s computations suggest that both procedures would do well with normal populations, while MM should outperform the standard method when there are outliers on Y (Anderson & Schumacker, 2003).

Slope tests for heterogeneous residual variances

Alternative parametric significance tests have been developed by Welch (1938), James (1951) and Deshon and Alexander (1994) to test for slope differences when residual variances are unequal. All three methods avoid the standard test’s pooling of groups’ residual variances in (6). Comparative research has shown that the three parametric alternative tests perform similarly in terms of robustness and power (Luh & Guo, 2000; Luh & Guo, 2002; Deshon & Alexander, 1996), so this study focuses solely on the James second-order test, which is slightly better than the Welch and Deshon and Alexander tests in terms of power and robustness to nonnormality.

The steps of the James second order test are as follows:

- 1) Define a James weight, w_{j_i} , based on each group slope’s standard error,

$$w_{j_i} = \frac{1/\sigma_{\beta_{j_i}}^2}{\sum_j 1/\sigma_{\beta_{j_i}}^2}. \quad (19)$$

STANDARD, ALTERNATIVE, AND ROBUST SLOPE TEST STRATEGIES

- 2) Define a variance-weighted common slope as,

$$\beta^+ = \sum_j w_{j,j} \beta_{1j} . \quad (20)$$

- 3) Define the James' test statistic as,

$$\text{James} = \sum_j \frac{(\beta_{1j} - \beta^+)^2}{\sigma_{\beta_{1j}}^2} . \quad (21)$$

- 4) Evaluate the significance of the James' test statistic by determining if it exceeds the following critical value,

$$\begin{aligned} \text{James}_{\text{crit}} = & c + (1/2)(3\chi_4 + \chi_2) \sum_j [(1 - w_{j,j})^2 / v_j] \\ & + (1/16)(3\chi_4 + \chi_2)^2 [1 - (J-3)/c] \sum_j [(1 - w_{j,j})^2 / v_j]^2 \\ & + (1/2)(3\chi_4 + \chi_2) [(8R_{23} - 10R_{22} + 4R_{21} - 6R_{12}^2 + 8R_{12}R_{11} \\ & - 4R_{11}^2) + (\chi_2 - 1)(2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} - 2R_{11}^2) \\ & + (1/4)(3\chi_4 - 2\chi_2 - 1)(4R_{12}R_{11} - R_{12}^2 - 2R_{12}R_{10} - 4R_{11}^2 \\ & + 4R_{11}R_{10} - R_{10}^2)] + (5\chi_6 + 2\chi_4 + \chi_2)(R_{23} - 3R_{22} + 3R_{21} - R_{20}) \\ & + (3/16)(35\chi_8 + 15\chi_6 + 9\chi_4 + 5\chi_2)(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} \\ & + R_{20}) + (1/16)(9\chi_8 - 3\chi_6 - 5\chi_4 - \chi_2)(4R_{21} - 2R_{22} - R_{20} + 2R_{12}R_{10} \\ & - 4R_{11}R_{10} + R_{10}^2) + (1/4)(27\chi_8 + 3\chi_6 + \chi_4 + \chi_2)(R_{11}^2 - R_{22}) \\ & + (1/4)(45\chi_8 + 9\chi_6 + 7\chi_4 + 3\chi_2)(R_{23} - R_{12}R_{11}) \end{aligned}$$

(22), where $v_j = N_j - 2$, c is the $1-\alpha$ quantile of the central chi-square distribution with $J-1$ degrees of freedom, $R_{ut} = \sum_j \frac{w_{j,j}^t}{v_j^u}$ and

$$\chi_{2s} = \frac{c^s}{\prod_{q=1}^s (J+2q-3)} \quad (\text{for } \chi_2, \chi_4, \chi_6, \text{ and } \chi_8,$$

s is 1, 2, 3, and 4, respectively).

Hybrid slope tests for nonnormal data and heterogeneous residual variances

Slope test strategies are not necessarily robust to problems for which they were not directly designed. The parametric alternative strategies that were designed to address residual variance heterogeneity have documented problems with nonnormal data (Deshon & Alexander, 1996). The slope test strategies that have been proposed for nonnormal data do not directly address residual variance heterogeneity.

An important area of research assesses so-called hybrid slope test strategies that may be robust to several assumption violations by use of nonnormality-robust group slopes and standard errors with parametric alternative tests that avoid the pooling of heterogeneous residual variances.

Recent research on hybrid slope test strategies has considered using standard slope estimates and standard errors or trimmed slope estimates and Winsorized standard errors with skew-corrected versions of parametric alternative tests (Luh & Guo, 2000; 2002). The use of the trimmed slopes and Winsorized standard errors with parametric alternative tests like James is straightforward, with groups' degrees of freedom calculated as $v_j = h_j - 1$ rather than as $N_j - 2$. Luh and Guo also transformed the test-statistics of the parametric alternatives to eliminate the effect of skewness (Johnson, 1978; Hall, 1992). For example, the proposed transformation for skewness for the James second order test statistic from (21) is, $\text{James_TT} =$

$$\sum_j \left[\left(\frac{\sqrt{N_j} [\frac{(\beta_{1j} - \beta^+)}{\sqrt{N_j} \sigma_{\beta_{1j}}} - \gamma_{x,j}^3 \gamma_{\varepsilon,j}^3 \frac{(\beta_{1j} - \beta^+)^2}{N_j \sigma_{\beta_{1j}}^2}]}{6 + \gamma_{x,j}^3 \gamma_{\varepsilon,j}^3 / (6N_j)} \right)^2 \right], \quad (23)$$

where $\gamma_{x,j}^3$ and $\gamma_{\varepsilon,j}^3$ are the sample skews of X and ε in group j . Luh and Guo's studies showed that their hybrid strategies were robust to both nonnormality and residual variance heterogeneity.

This study

This study extends prior research on the relative performance of slope testing strategies under nonnormality and residual variance heterogeneity. This study directly compares the standard, MM, and Theil-Sen tests, extending the previous comparisons based on estimating one slope that have given recommendations for MM regression over the standard method (Anderson & Schumaker, 2003) and for Theil-Sen over MM regression and the standard method (Wilcox & Keselman, 2004). The comparison of the trimmed and Winsorized slope test with the Theil-Sen and MM methods

has not been considered in previous studies, and it allows for an evaluation of some trimming (trimmed and Winsorized) with the most extreme trimming possible (Theil-Sen).

This study also extends Luh and Guo's (2000, 2002) work, first by separately evaluating the trimmed and Winsorized slope test and the skewness transformation of the James test statistic. Because the accuracy of slope estimation has more to do with the heaviness of the distribution's tails rather than its skew (Klockars & Moses, 2002), the test statistic transformation ought to have a smaller impact in correcting for nonnormality than the trimmed and Winsorized, Theil-Sen and MM methods. Finally, Luh and Guo's efforts to form hybrid slope test strategies that are robust to both nonnormality and residual variance heterogeneity are extended to consider hybrid slope tests based not only on integrating the trimmed and Winsorized methods and the skewness transformation with James second order method, but also the MM and Theil-Sen methods.

Methodology

A simulation study was conducted to investigate the relative robustness and power of the slope test strategies for comparing two groups' slopes. Empirical rejection rates of the null hypothesis were computed based on 10,000 replications for each condition. Two treatment groups were used throughout the study. The following conditions were considered.

Slope Test Strategies

Five stand-alone slope test strategies and four hybrids of the five strategies were evaluated.

- 1) The standard F test of slope differences in (6) (Standard).
- 2) The James parametric alternative test in (21) (James).
- 3) Significance testing of the β_3 in model (2) based on MM estimation with the default settings in SAS PROC ROBUSTREG (SAS Institute Inc., 2003) (MM).
- 4) The trimmed and Winsorized slope test in (11) using 10% trimming (TW).

- 5) The Theil-Sen estimator with percentile bootstrapping for the significance testing (TS).

The following four hybrid strategies were also considered:

- 6) The James procedure with the Johnson's one-sample t-statistic transformation for skewness in (23) (James-TT).
- 7) The James procedure using MM slope estimates and standard errors (James-MM).
- 8) The James procedure using 10% trimmed slope estimates and Winsorized standard errors from Luh and Guo (2000) (James-TW).
- 9) The James procedure using the Theil-Sen slope estimates and the standard deviations of 599 bootstrapped Theil-Sen estimates from strategy 5 for the group slopes' standard errors (James-TS).

Defining groups' observations and degrees of freedom

For the James-MM and James-TS strategies, some consideration was given for defining the groups' degrees of freedom. Initial efforts were based on Luh and Guo's (2000) attempt to account for the number of observations used in the slope estimate in James-TW ($v_j = h_j - 1$). Directly applying this to James-TS would mean setting $v_j = 2 - 1$. From initial results it was clear that using $v_j = 1$ resulted in extremely conservative tests for James-TS, so in an effort to obtain more reasonable results, the v_j was set as $N_j(N_j - 1)/2 - 1$. For James-MM, degrees of freedom were set to account for the weighting of the observations used in the MM slope estimate,

$$v_j = \sum_{i \in j} \frac{\partial \xi(\frac{\epsilon_i}{\sigma})}{\partial \left(\frac{\epsilon_i}{\sigma}\right)} \frac{1}{\left(\frac{\epsilon_i}{\sigma}\right)} - 2. \quad \text{This } v_j \text{ produced}$$

James-MM results that were very similar to setting $v_j = N_j - 2$.

STANDARD, ALTERNATIVE, AND ROBUST SLOPE TEST STRATEGIES

Y's Distribution

Eight shapes were used for Y, including a normal shape (skew=0, kurtosis=0), and seven other shapes with various degrees of skews and kurtosis (Table 1).

Variance heterogeneity

The two considered residual variance ratios for the groups were 1/1 and 3/1. For conditions of unequal sample size, the residual variances were directly and inversely paired with the treatment group sample sizes.

Sample sizes

Twenty and forty subjects per treatment group were used. The conditions of unequal sample size used twenty subjects in one group and forty in the other.

Data generation method: Robustness

The following data generation method was used to create X and Y variables of desired distributions and variances with equal slopes in the two groups.

- 1) N values of one standard normal variate, Z, were generated, where N is the total sample size in two groups.
- 2) Y was created as a transformation of Z using Fleishman's (1978) method for generating nonnormal variables:

$$Y = a + bZ + cZ^2 + dZ^3 \quad (24)$$

The constants (a, b, c, and d) and resulting distributions are listed in Table 1.

- 3) An error variable for X (ε) was generated as a standard normal variate. X's degree of nonnormality was a compromise between Y's nonnormality and ε 's normality.
- 4) Desired numbers of Ys and ε s were randomly assigned to treatment groups 1 and 2.
- 5) X was created as a function of Y and ε :

$$X_{ij} = \rho_j Y_{ij} + \sqrt{(1 - \rho_j^2)} \varepsilon_{ij}, \quad (25)$$

where ρ_j is the desired XY correlation for treatment group j.

- 5) Y_{ij} was multiplied by a number, σ_{Yj} , that resulted in a desired standard

deviation for Y in the jth treatment group and, in conjunction with ρ_j , a desired residual variance. The values of σ_{Yj} and ρ_j for the two groups achieved a particular residual variance ratio (Table 2), while keeping the slopes equal in the two groups.

Data generation method: Power

The data generation process used to assess strategies' power was similar to the data generation process used to assess robustness. All variables' distributions were normal. One group's XY correlation and Y standard deviation were 0.5 and 1.0, respectively, while the second group's XY correlation and Y standard deviation were 0.0 and 0.866, respectively. The XY correlations and Y standard deviations across the groups resulted in a population slope difference of 0.5 while meeting the normality and equal residual variances assumptions of the standard test.

Analysis strategy

The assessment of strategies' robustness involved comparing their average rejection rates to the nominal 0.05 rate for conditions where no slope differences existed in the population. Deviations from the nominal 0.05 rate were determined to be excessively conservative or liberal when they were outside of two standard errors band reflective of the number of replications used in this study

$$(0.05 + /- 2 \sqrt{\frac{(.05)(.95)}{10,000}} = 0.046 \text{ to } 0.054). \quad \text{The}$$

standard error band roughly corresponded to Bradley's (1978) conservative range for robust Type I error rates, 0.045 to 0.055.

The assessment of strategies' power involved comparing strategies' average rejection rates to each other for conditions where actual slope differences existed in the population.

Follow-up analyses were also conducted to gain further insight into how the slope estimation strategies were working in the conditions of this study. These follow-up analyses included assessments of averages and standard deviations of the strategies' slope estimates to indicate their bias and efficiency,

and assessments of strategies' average standard errors to provide understanding of the accuracy of strategies' significance tests.

Results

Tables 3-9 present the considered strategies' empirical Type I error rates across the 56 combinations of nonnormality, residual variance heterogeneity and sample size. Nonnormality affected the Standard, James and MM tests similarly, creating liberal Type I error rates when the Y distributions were leptokurtic and conservative Type I error rates when the distributions were platykurtic. The TW test had Type I error rates that were close to the nominal rate across the conditions of nonnormality. The TS test had Type I error rates that were consistently conservative across the considered levels of nonnormality. In terms of the hybrid strategies, James-TT had Type I error rates that were almost indistinguishable from James, while the James-MM, James-TW and James-TS strategies had Type I error rates reflective of the nonnormality strategy used, being excessively liberal for James-MM, being near 0.05 for James-TW, and being excessively conservative for James-TS.

The effect of residual variance heterogeneity on Type I error differed for the equal and unequal sample size conditions. When sample sizes were equal (Tables 4 & 9), MM was the only strategy affected by residual variance heterogeneity, becoming excessively liberal. When sample sizes were unequal (Tables 6 & 7), the groups' sample size-residual variance pairing affected the Standard, MM and TW tests similarly, making them liberal with an inverse pairing and conservative with a direct pairing. The James hybrid strategies were largely unaffected by the combination of unequal sample sizes and residual variances. James-TS produced conservative Type I error rates for most of the considered residual variance conditions.

The effect of combining nonnormality and residual variance heterogeneity (Tables 4, 6, 7, & 9) produced somewhat unique Type I error patterns for the nine tests. For the Standard test, residual variance heterogeneity usually made the effect of nonnormality less extreme except for when sample sizes were inversely-paired with

residual variances, in which case Type I error was made more extreme. For James and James-TT, residual variance heterogeneity made the effects of nonnormality less extreme, though James did not react as much to the combination of unequal sample sizes and residual variances as the Standard test. The MM test often had the most problematic Type I error rates for combinations of nonnormality and residual variance heterogeneity. The TW and TS tests were not particularly affected by the combination of nonnormality and residual variance heterogeneity, where the TW strategy was mainly impacted by the combination of unequal sample sizes and residual variances while the TS strategy was largely uninfluenced by anything. The Type I errors of hybrid strategies were reflective of the nonnormality strategy on which they were based, being liberal for James-MM, conservative for James-TS, and close to the 0.05 level for James-TW.

Power

Table 10 compares the power of the nine strategies across three considered sample size conditions with normal distributions, equal residual variances and a population slope difference of 0.5. The most powerful strategies were the Standard, James and James-TT strategies, of which there was no overwhelming winner. The MM test had lower power rates than the Standard, James and James-TT tests. The James-MM hybrid strategy had less power than the MM strategy. The TW and James-TW tests had the lowest power rates of the considered strategies. The James-TS and TS strategies had higher power rates than the TW and James-TW strategies and (mostly) lower power rates than the MM and James-MM strategies. The use of TS as a hybrid with James (James-TS) increased its power relative to the TS strategy.

Slope Estimation

To gain further insight into the four slope estimation methods (Standard, MM, TW and TS), Table 11 summarizes each methods' 10,000 estimates of one slope with population value 0.5 in samples of size 20. When distributions were normal, all four methods gave average slope values close to 0.5. The methods' standard deviations show that the Standard

method's estimates were least variable, followed by the MM estimates, the TS estimates and finally the TW estimates (corresponding to TW's relatively low power). The methods' average estimated standard errors correspond to the overall liberalness/conservativeness of the methods' significance tests, and for normal distributions show that on average all methods except for TS have standard errors that closely approximate slope variability. TS's bootstrapped standard errors over-estimated TS slope variability, corresponding to the conservativeness of its Type I error rates.

The slope estimation results in Table 11 for a leptokurtic Y (kurtosis = 12) differ from those for a normal Y (kurtosis = 0). For a leptokurtic Y, all estimation methods underestimate the population slope value of 0.5, where the least biased estimator is the Standard method while the most biased is the MM estimate. The Standard method's slope estimates are the most variable while the TS estimates are the least variable. The average standard errors of the Standard test and MM underestimate slope variability, corresponding to the liberalness of the Standard's and MM's Type I error rates. The TW estimates have standard errors that slightly underestimate slope variability. The TS estimate has standard errors that overestimate slope variability, corresponding to the conservativeness of TS. The results in Table 11 support previous findings that the TS estimator is more stable than the MM and Standard estimates when distributions are nonnormal (Wilcox & Keselman, 2004). These results extend previous work by showing that with nonnormality, the Standard test and MM regression underestimate slope variability (making the Type I error rates of the Standard and MM slope tests liberal), the Winsorized standard errors provide relatively accurate estimates of the variability of the trimmed slopes, while the TS bootstrap method overestimates slope variability (making the Type I error rates of the TS slope test conservative).

Conclusion

The purpose of this study was to compare some recently-researched strategies for testing independent groups' regression slopes. The standard test of slope differences was shown to

have its usual robustness problems with respect to nonnormality and the pairing of unequal sample sizes and residual variances. Alternative strategies proposed for addressing nonnormality and used in hybrid strategies for addressing both nonnormality and residual variance heterogeneity were also assessed. The most promising of the alternative strategies in terms of robustness and power were the Theil-Sen strategy and a hybrid of Theil-Sen and the James second-order parametric alternative test. These Theil-Sen strategies had somewhat conservative Type I error rates that were largely unaffected by nonnormality and residual variance heterogeneity, and slope estimates that were efficient even for nonnormal data. The hybrid strategy of trimming and Winsorizing slope estimates and using them with the James test had Type I error rates that were closest of all the considered strategies to the nominal 0.05 level, but trimming and Winsorizing also produced slope tests with the lowest power rates of the considered strategies. Of the other strategies considered, James, James with a test statistic transformation for skewness, MM regression and the use of MM estimates with James are not recommended due to their robustness problems with nonnormal data.

In evaluating the results of this and other studies, it is important to acknowledge that the effects of nonnormality have been considered in very different ways, all of which have implications for studies' results. When the nonnormality of ε is directly manipulated, the standard and James tests have appeared to be robust to all but the most extreme shapes (e.g., skew=6.2, kurtosis=114 in Luh & Guo, 2000, 2002).

When the nonnormality of Y and/or X is manipulated, the standard and James tests become problematic for relatively small degrees of nonnormality (e.g., skew=1.95, kurtosis=7.69 in Deshon & Alexander, 1996). When nonnormality has been studied in terms of outliers in multivariate distributions, the standard test is problematic and MM regression performs well (Anderson & Schumacker, 2003). The second type of nonnormality, in Y and X, creates great problems for methods that use least squares estimation methods due to the higher likelihood of leverage points.

MOSES & KLOCKARS

Table 1. Shapes and Fleishman constants used to generate the variables

Skew	Kurtosis	a	b	c (= -a)	d
0	-1.15	0	1.34	0	-0.132
0	0	0	1	0	0
1.2	1.11	-0.340774	1.095718	0.340774	-0.080735

Table 2. Correlations and standard deviations used to create levels of residual variance heterogeneity.

Residual Variance Ratio	ρ_1	σ_{Y1}	ρ_2	σ_{Y2}
1/1	0.5	1.0	0.5	1.0
1/3	0.5	1.0	0.3162	1.5811

Table 3. Empirical Type I error rates for group sample sizes of 20, 20 and a residual variance ratio of 1/1.

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James- TT	James- MM	James- TW	James- TS
0	-1.15	0.0256*	0.0257*	0.0260*	0.0479	0.0207*	0.0258*	0.0214*	0.0456	0.0204*
0	0	0.0486	0.0481	0.0610*	0.0497	0.0282*	0.0483	0.0486	0.0479	0.0262*
1.2	1.11	0.0668*	0.0676*	0.1099*	0.0508	0.0304*	0.0667*	0.0985*	0.0490	0.0358*
1.6	2.86	0.0941*	0.0961*	0.1383*	0.046*	0.0302*	0.0949*	0.1341*	0.0455*	0.0345*
0	3	0.0912*	0.0936*	0.0999*	0.0532	0.0307*	0.0935*	0.0874*	0.0508	0.0317*
0	6	0.1178*	0.1200*	0.1226*	0.0500	0.0341*	0.1198*	0.1091*	0.0482	0.0328*
0	9	0.1359*	0.1403*	0.1257*	0.0458	0.0308*	0.1395*	0.1193*	0.0429*	0.0288*
0	12	0.1645*	0.1727*	0.1347*	0.0542	0.0303*	0.1714*	0.1343*	0.0510	0.0281*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

Table 4. Empirical Type I error rates for group sample sizes of 20, 20 and a residual variance ratio of 3/1.

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James- TT	James- MM	James- TW	James- TS
0	-1.15	0.0428*	0.0398*	0.0820*	0.0566*	0.0291*	0.0393*	0.0346*	0.0531	0.0281*
0	0	0.0514	0.0481	0.1020*	0.0511	0.0323*	0.0482	0.0438*	0.0486	0.0305*
1.2	1.11	0.0630*	0.0598*	0.1339*	0.0550*	0.0304*	0.0591*	0.0809*	0.0509	0.0292*
1.6	2.86	0.0792*	0.0774*	0.1525*	0.0517	0.0317*	0.0757*	0.1063*	0.0477	0.0278*
0	3	0.0689*	0.0682*	0.1141*	0.0523	0.0303*	0.0674*	0.0653*	0.0480	0.0302*
0	6	0.0931*	0.0946*	0.1286*	0.0517	0.0304*	0.0938*	0.0824*	0.0470	0.0251*
0	9	0.1106*	0.1096*	0.1368*	0.0481	0.0326*	0.1091*	0.0886*	0.0440*	0.0263*
0	12	0.1176*	0.1206*	0.1367*	0.0479	0.0323*	0.1202*	0.0904*	0.0437*	0.0235*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

STANDARD, ALTERNATIVE, AND ROBUST SLOPE TEST STRATEGIES

Table 5. Empirical Type I error rates for group sample sizes of 20, 40 and a residual variance ratio of 1/1.

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James-TT	James-MM	James-TW	James-TS
0	-1.15	0.0300*	0.0306*	0.0247*	0.0546*	0.0263*	0.0304*	0.0252*	0.0532	0.0259*
0	0	0.0496	0.0486	0.0581	0.0507	0.0314*	0.0482	0.0506	0.0468	0.0320*
1.2	1.11	0.0701*	0.0720*	0.1174*	0.0571*	0.0327*	0.0713*	0.1090*	0.0504	0.0407*
1.6	2.86	0.1040*	0.1054*	0.1451*	0.0556*	0.0321*	0.1049*	0.1530*	0.0526	0.0412*
0	3	0.0931*	0.0929*	0.0981*	0.0531	0.0327*	0.0927*	0.0870*	0.0482	0.0349*
0	6	0.1235*	0.1286*	0.1143*	0.0525	0.0325*	0.1269*	0.1080*	0.0482	0.0341*
0	9	0.1524*	0.1599*	0.1226*	0.0522	0.0371*	0.1593*	0.1225*	0.0485	0.0356*
0	12	0.1677*	0.1804*	0.1253*	0.0528	0.0349*	0.1799*	0.1293*	0.0508	0.0338*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

Table 6. Empirical Type I error rates for group sample sizes of 20, 40 and a residual variance ratio of 1/3 (Direct Pairing).

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James-TT	James-MM	James-TW	James-TS
0	-1.15	0.0134*	0.0353*	0.0251*	0.0273*	0.0312*	0.0352*	0.0299*	0.0492	0.0295*
0	0	0.0218*	0.0534	0.0406*	0.0258*	0.0313*	0.0535	0.0508	0.0505	0.0335*
1.2	1.11	0.0324*	0.0660*	0.0851*	0.0274*	0.0279*	0.0656*	0.0913*	0.0499	0.0358*
1.6	2.86	0.0443*	0.0905*	0.1171*	0.0269*	0.0345*	0.0893*	0.1203*	0.0488	0.0383*
0	3	0.0379*	0.0789*	0.0654*	0.0244*	0.0349*	0.0792*	0.0746*	0.0488	0.0352*
0	6	0.0612*	0.1087*	0.0883*	0.0283*	0.0348*	0.1088*	0.0953*	0.0493	0.0363*
0	9	0.0765*	0.1327*	0.0997*	0.0290*	0.0347*	0.1319*	0.1011*	0.0466	0.0344*
0	12	0.0900*	0.1516*	0.1172*	0.0294*	0.0342*	0.1503*	0.1151*	0.0486	0.0286*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

Table 7. Empirical Type I error rates for group sample sizes of 40, 20 and a residual variance ratio of 1/3 (Inverse Pairing).

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James-TT	James-MM	James-TW	James-TS
0	-1.15	0.0827*	0.0360*	0.1722*	0.1028*	0.0305*	0.0356*	0.0342*	0.0497	0.0337*
0	0	0.0986*	0.0504	0.1860*	0.0989*	0.0341*	0.0504	0.0497	0.0477	0.0320*
1.2	1.11	0.1165*	0.0624*	0.2039*	0.0981*	0.0347*	0.0615*	0.0860*	0.0509	0.0342*
1.6	2.86	0.1392*	0.0798*	0.2017*	0.0954*	0.0386*	0.0769*	0.0998*	0.0503	0.0312*
0	3	0.1359*	0.0760*	0.1880*	0.1007*	0.0370*	0.0760*	0.0682*	0.0510	0.0343*
0	6	0.1620*	0.0997*	0.1902*	0.1032*	0.0354*	0.0995*	0.0797*	0.0531	0.0311*
0	9	0.1812*	0.1164*	0.1879*	0.0936*	0.0390*	0.1148*	0.0863*	0.0467	0.0301*
0	12	0.1862*	0.1263*	0.1810*	0.0887*	0.0391*	0.1260*	0.0903*	0.0471	0.0291*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

MOSES & KLOCKARS

Table 8. Empirical Type I error rates for group sample sizes of 40, 40 and a residual variance ratio of 1/1.

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James-TT	James-MM	James-TW	James-TS
0	-1.15	0.0280*	0.0284*	0.0206*	0.0526	0.0289*	0.0283*	0.0200*	0.0526	0.0300*
0	0	0.0479	0.0477	0.0559*	0.0504	0.0321*	0.0479	0.0498	0.0501	0.0353*
1.2	1.11	0.0672*	0.0677*	0.1115*	0.0484	0.0388*	0.0678*	0.1101*	0.0478	0.0521
1.6	2.86	0.1028*	0.1042*	0.1435*	0.0495	0.0371*	0.1032*	0.1635*	0.0494	0.0483
0	3	0.1006*	0.1028*	0.0900*	0.0518	0.0391*	0.1030*	0.0878*	0.0518	0.0428*
0	6	0.1396*	0.1436*	0.1044*	0.0489	0.0352*	0.1432*	0.1057*	0.0483	0.0381*
0	9	0.1709*	0.1752*	0.1151*	0.0467	0.0389*	0.1743*	0.1235*	0.0458	0.0398*
0	12	0.1952*	0.2004*	0.1200*	0.0483	0.0396*	0.1990*	0.1347*	0.0476	0.0387*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

Table 9. Empirical Type I error rates for group sample sizes of 40, 40 and a residual variance ratio of 3/1.

Skew	Kurtosis	Standard	James	MM	TW	TS	Hybrid Strategies			
							James-TT	James-MM	James-TW	James-TS
0	-1.15	0.0360*	0.0354*	0.0809*	0.0506	0.0352*	0.0353*	0.0310*	0.0494	0.0388*
0	0	0.0494	0.0479	0.0884*	0.0506	0.0386*	0.0482	0.0470	0.0490	0.0423*
1.2	1.11	0.0662*	0.0636*	0.1511*	0.0499	0.0432*	0.0639*	0.0958*	0.0484	0.0503
1.6	2.86	0.0851*	0.0849*	0.1828*	0.0515	0.0390*	0.0841*	0.1236*	0.0506	0.0434*
0	3	0.0844*	0.0821*	0.1191*	0.0548	0.0386*	0.0820*	0.0733*	0.0520	0.0400*
0	6	0.1025*	0.1011*	0.1279*	0.0528	0.0386*	0.1008*	0.0820*	0.0513	0.0378*
0	9	0.1272*	0.1290*	0.1427*	0.0512	0.0408*	0.1278*	0.0945*	0.0482	0.0386*
0	12	0.1426*	0.1460*	0.1431*	0.0486	0.0367*	0.1439*	0.0998*	0.0474	0.0322*

* Type I error rates outside +/- 2 standard errors of the nominal 0.0500 rate (0.0456 to 0.0544).

Table 10. Empirical Power rates for population slope differences of .5 and normality and residual variance assumptions met.

Sample Sizes		Standard	James	MM	TW	TS	Hybrid Strategies			
Group 1	Group 2						James-TT	James-MM	James-TW	James-TS
20	20	0.3910	0.3901	0.3735	0.2270	0.2655	0.3906	0.3372	0.2215	0.2891
20	40	0.5096	0.4989	0.4730	0.3082	0.3931	0.4994	0.4403	0.2930	0.4202
40	40	0.6909	0.6912	0.6359	0.4369	0.6179	0.6912	0.6180	0.4353	0.6483

STANDARD, ALTERNATIVE, AND ROBUST SLOPE TEST STRATEGIES

Table 11. Descriptive analyses for the slope estimation strategies (population slope = 0.5).

Skew	Kurtosis	Sample Size		Standard	MM	TW	TS
0	0	20	Mean of Slope Estimates	0.5043	0.5043	0.5079	0.4998
			Standard Deviation of Slope Estimates	0.2080	0.2214	0.2819	0.2275
			Mean of Slope Standard Error Estimates	0.2046	0.2205	0.2721	0.2752
0	12	20	Mean of Slope Estimates	0.4598	0.2987	0.3538	0.3280
			Standard Deviation of Slope Estimates	0.2866	0.2361	0.2528	0.2159
			Mean of Slope Standard Error Estimates	0.1907	0.1701	0.2383	0.2575

Figure 1

Figure 2

Figure 3

Figure 4

This type of nonnormality was of most interest in this study. It caused serious problems for the standard test that warranted the use of alternative and robust strategies, and it created data situations that differentiated all of the studied strategies.

To gain some final insight into the for considered slope estimation methods, a representative sample of twenty observations was generated from this study's kurtosis=12 condition. Figure 1 shows these XY data and plots the Standard, least squares regression line. There is one very extreme X observation (almost 3 standard deviations from X's population mean of zero) that is also very low on Y (i.e., a bad leverage point). This observation causes the standard slope estimation method to underestimate the population slope of 0.5 in its slope estimate of 0.421. Figures 2 and 3 plot the observations in the data that are not excluded in computing the trimmed slope (Figure 2) and the Theil-Sen slope (Figure 3). The trimmed and Theil-Sen methods underestimate the population slope more than the Standard method, producing slope estimates of 0.393 and 0.231, respectively.

Figure 4 is especially useful for understanding the very complicated MM regression procedure. All twenty of the original observations are used in MM regression, but contribute in weighted form to the final MM slope estimate. The observations' weights in Figure 4 show that the high-leverage observation is weighed very heavily by the MM method, causing the MM slope estimate to be relatively small (0.148). The observations that are far from the MM regression line are assigned small weights. Figure 4 shows that with MM high-leverage points can be weighted such that they influence the final slope estimate much more than the Standard least squares estimate. The large weights that are assigned to high leverage points in MM result in MM standard errors that underestimate slope variability (the W in equation 18 is large) and inflate the Type I error of the MM strategy. Figure 4 makes it clear that the problems of the MM strategy with respect to high leverage points are not likely to be fixed by altering the weighting function, ξ , or the κ that determines how each of the scaled residuals are weighted. It may be possible to address MM's

problems with high leverage data points through a wise choice of starting values that define the MM regression line and the residuals with respect to this line.

Implications

This article considered some of the recently-researched slope test strategies. Some of the strategies not considered in this paper were excluded because they have had noted problems and criticisms, including nonparametric alternative tests (Marascuilo, 1966; Dretzke, Levin & Serlin, 1982; Deshon & Alexander, 1996), residuals-based bootstrapping (Luh & Guo, 2000), ranked data (Headrick & Sawilowsky, 2000; Klockars & Moses, 2002), data transformations (Wilcox & Keselman, 2004; Aguinis & Pierce, 1998; Keselman, Carriere & Lix, 1995; Glass, Peckham & Saunders, 1972), several robust regression strategies (Anderson & Schumaker, 2003) and judgment-based elimination of outliers (Wilcox, 1996; He & Portnoy, 1992).

There are other strategies that are variations on the ones considered in this study, such as the use of Theil-Sen after trimming outliers (Wilcox & Keselman, 2004), the use of Theil-Sen based on less than the $N(N-1)/2$ slopes that could be created out of all pairs of observations (Ebrahim & Al-Nasser, 2005), other parametric alternative tests for residual variance heterogeneity (Alexander & Deshon, 1994; Welch, 1938), and trimmed and Winsorized estimates with varied amounts of trimming.

The results of this study suggest that an especially promising slope test strategy would combine the best features of the trimming and Winsorizing methods with Theil-Sen. By using the trimming and Winsorizing strategy on the $N(N-1)/2$ slopes that could be created out of all pairs of observations rather than only $N/2$ pairs, the final trimmed slope estimates should have stability levels that are similar to those of Theil-Sen, ultimately improving the power of the trimmed and Winsorized slope test. This proposed test would avoid the excessively time consuming and excessively-conservative bootstrapping that accompanies the Theil-Sen method, reduce the bias of the Theil-Sen estimates for nonnormal data, provide a

reasonable answer to the awkward definition of the number of observations used by the median-based Theil-Sen, and provide the analyst some flexibility in terms of the extent of trimming used in the final slope estimates. A study that considers how the number of slopes (Ebrahem & Al-Nasser, 2005) and the extent of trimming contribute to Type I error and power across conditions of nonnormality would be especially useful for creating the best version of this proposed test of slope differences.

References

- Aguinis, H., & Pierce, C. A. (1998). Heterogeneity of error variance and the assessment of moderating effects of categorical variables: A conceptual review. *Organizational Research Methods, 1*, 296-314.
- Aguinis, H., Petersen, S. A., & Pierce, C. A. (1999). Appraisal of the homogeneity of error variance assumption and alternatives to multiple regression for estimating moderating effects of categorical variables. *Organizational Research Methods, 2*(4), 315-339.
- Aiken, L. S. & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.
- Alexander, R. A., & DeShon, R. P. (1994). Effect of error variance heterogeneity on the power of tests for regression slope differences. *Psychological Bulletin, 115*(2), 308-314.
- Anderson, C., & Schumacker, R. E. (2003). A comparison of five robust regression methods with ordinary least squares regression: Relative efficiency, bias, and test of the null hypothesis. *Understanding Statistics, 2*(2), 79-103.
- Bradley, J. C. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.
- Conerly, M. D. & Mansfield, E. R. (1988). An approximate test for comparing heteroscedastic regression models. *Journal of the American Statistical Association, 83*(403), 811-817.
- Conover, W. J. & Iman, R. L. (1982). Analysis of covariance using the rank transformation. *Biometrics, 38*, 715-724.
- Cronbach, L. J. & Snow, R. E. (1977). *Aptitude and instructional methods*. Irvington, NY.
- Dance, K. A., & Neufeld, R. W. J. (1988). Aptitude-treatment interaction research in the clinical setting: A review of attempts to dispel the "Patient Uniformity" myth. *Psychological Bulletin, 104*(2), 192-213.
- Deshon, R. P., & Alexander, R. A. (1994). A generalization of James' second-order approximation to the test for regression slope equality. *Educational and Psychological Measurement, 54*, 328-335.
- DeShon, R. P. & Alexander, R. A. (1996). Alternative procedures for testing regression slope homogeneity when group error variances are unequal. *Psychological Methods, 1*(3), 261-277.
- Dretzke, B. J. & Levin, J. R.; Serlin, R. C. (1982). Testing for regression homogeneity under variance heterogeneity. *Psychological Bulletin, 91*, 376-383.
- Ebrahem, M. A., & Al-Nasser, A. D. (2005). Estimating the slope of simple linear regression in the presence of outliers. *Journal of Modern Applied Statistical Methods, 4*(2), 509-513.
- Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532.
- Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research, 42*, 237-288.
- Guo, J. H. (1996). Distribution-free test for slope in linear regression model with replications. *Communications in Statistics-Theory and Methods, 25*, 2069-2078.
- Hall, P. (1992). On the removal of skewness by transformation. *Journal of Royal Statistical Society, B54*, 221-228.
- He, X., & Portnoy, S. (1992). Reweighted LS estimators converge at the same rate as the initial estimator. *Annals of Statistics, 20*, 2161-2167.
- Headrick, T. C. & Sawilowsky, S. S. (2000). Properties of the rank transformation in factorial analysis of covariance. *Communications in Statistics- Simulation and Computation, 29*, 1059-1087.

- Holland, P. & Welsch, R. (1977). Robust regression using interactively reweighted least squares. *Communications in Statistics-Theor. Meth.*, 6, 813-827.
- Hunter, J. E., Schmidt, F. L., & Hunter, R. (1979). Differential validity of employment tests by race: A comprehensive review and analysis. *Psychological Bulletin*, 86(4), 721-735.
- James, G. S. (1951). The comparison of several groups of observations when the ratios of population variances is unknown. *Biometrika*, 38, 324-329.
- Johnson, N. J. (1978). Modified *t* tests and confidence intervals for asymmetrical populations. *Journal of the American Statistical Association*, 73, 536-544.
- Keselman, H. J., Carriere, K. C., & Lix, L. M. (1995). Robust and powerful nonorthogonal analyses. *Psychometrika*, 60(3), 395-418.
- 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.
- Klockars, A. & Moses, T. (2002). Type I error rates for rank-based tests of homogeneity of slopes. *Journal of Modern Applied Statistical Methods*, 1(2), 452-460.
- Luh, W., & Guo, J. (2000). Approximate transformation trimmed mean methods to the test of simple linear regression slope equality. *Journal of Applied Statistics*, 27(7), 843-857.
- Luh, W., & Guo, J. (2002). Using Johnson's transformation with approximate test statistics for the simple regression slope homogeneity. *The Journal of Experimental Education*, 71(1), 69-81.
- Marascuilo, L. A. (1966). Large-sample multiple comparisons. *Psychological Bulletin*, 65, 280-290.
- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105, 156-166.
- Moses, T., & Klockars, A. (2005). Testing for aptitude-treatment interactions in analysis of covariance and randomized block designs under assumption violations. *Journal of Modern Applied Statistical Methods*, 4(2), 460-472.
- Myers, J. L. & Well, A. D. (1995). *Research design and statistical analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Overton, R. C. (2001). Moderated multiple regression for interactions involving categorical variables: A statistical control for heterogeneous variance across two groups. *Psychological Methods*, 6(3), 218-233.
- Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Fort Worth, Texas.
- Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P. J. & Van Driessen, K. (2000). An algorithm for positive-breakdown regression based on concentration steps. *Data Analysis: Scientific Modeling and Practical Application*, ed. W. Gaul, O. Optiz, & M. Schader. New York: Springer-Verlag, 335-346.
- SAS Institute Inc. (2003). *SAS/STAT user's guide, version 9.1: The ROBUSTREG Procedure*. Cary, NC: SAS Institute.
- Seligman, M. E. P. (2002). *Authentic happiness: Using the new positive psychology to realize your potential for lasting fulfillment*. Free Press, New York.
- Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, 63, 1379-1389.
- Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae*, 12, 85-91.
- Wang, X. (2005). Asymptotics of the Theil-Sen estimator in the simple linear regression model with a random covariate. *Nonparametric Statistics*, 17(1), 107-120.
- Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29, 350-362.
- Wilcox, R. R. (1996). A review of some recent developments in robust regression. *British Journal of Mathematical and Statistical Psychology*, 49, 253-274.
- Wilcox, R. R. (2004). Some results on extensions and modifications of the Theil-Sen regression estimator. *British Journal of Mathematical and Statistical Psychology*, 57, 265-280.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
- Wilcox, R. R., & Keselman, H. J. (2004). Robust regression methods: Achieving small standard errors when there is heteroscedasticity. *Understanding Statistics*, 3(4), 349-364.
- Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics*, 15, 642-656.

Selection of Non-Regular Fractional Factorial Designs When Some Two-Factor Interactions are Important

Weiming Ke

South Dakota State University

Rui Yao

University of Missouri-Columbia

A new method is proposed for selecting the optimal non-regular fractional factorial designs in the situation when some two-factor interactions are potentially important. Searching for the best designs according to this method is discussed and some results for the Plackett-Burman design of 12 runs are presented.

Key words: Alias matrix, fractional factorial design, non-regular design, partial confounding, Plackett-Burman design.

Introduction

Non-regular two-level fractional factorial designs, such as Plackett–Burman designs, are becoming popular choices in many areas of scientific investigation due to their run size economy and flexibility. The run size of non-regular two-level factorial designs is a multiple of 4. They fill the gaps left by the regular two-level fractional factorial designs whose run size is always a power of 2 (4, 8, 16, 32, ...). In non-regular factorial designs each main effect is partially confounded with all the two-factor interactions not involving itself. Because of this complex aliasing structure, non-regular factorial designs had not received sufficient attention

until recently. A number of authors studied the projection properties of two-level non-regular factorial designs. This includes Box and Bisgaard (1993), Lin and Draper (1993, 1995), Cheng (1995, 1998), Box and Tyssedal (1996), and Dean and Draper (1999). More recently, Deng and Tang (1999) proposed generalized resolution and minimum aberration criteria for ranking non-regular two-level factorial designs in a systematic way. Their criteria were further studied by Tang and Deng (1999), Tang (2001), Xu and Wu (2001), Ma and Fang (2001), and Butler (2003). Based on the generalized minimum aberration criteria, Deng, Li, and Tang (2000) and Deng and Tang (2002) provided tables of non-regular designs with favorable aberration properties for $n \leq 24$ runs, Cheng, Li and Ye (2004) studied optimal blocking schemes for non-regular designs. Despite the above important contributions, a basic problem in this area still remains unsolved. The problem is how to assess, compare, and rank non-regular factorial designs when some two-factor interactions are potentially important.

In practical applications of non-regular designs, it is often in the case that some of the two-factor interactions are important and need to be estimated in addition to the main effects. In this article, we consider how to select non-regular two-level fractional factorial designs when some of the two-factor interactions are presumably important. We propose and study a method to select the optimal non-regular two-level fractional factorial designs in the situation that some of the two-factor interactions are

Weiming Ke is an Assistant Professor of Statistics in the Department of Mathematics and Statistics, South Dakota State University. Research interests: design of experiments, survival analysis, stochastic models, and biostatistics. Email: weiming.ke@sdstate.edu. Rui Yao is an Assistant Professor in the Department of Personal Financial Planning, University of Missouri-Columbia, and was formerly at South Dakota State University. Research interests: applied statistics, individual financial risk tolerance, financial ratios, portfolio allocation, debt management, retirement preparation, and household consumption patterns. Email: yaor@missouri.edu.

potentially important. We then discuss how to search for the best designs according to this method and present some results for the Plackett-Burman design of 12 runs.

Non-regular fractional factorial designs

Non-regular fractional factorial designs are commonly obtained from Plackett-Burman designs or Hadamard matrices in general by selecting a subset of the columns. Plackett and Burman (1946) provided a series of two-level fractional factorial designs for examining $(n - 1)$ factors in n run, where n is a multiple of 4. For example, a 12-run Plackett-Burman design can be constructed by shifting the row $(+ + - + + - - - + -)$ one place to the right 10 times and then adding a vector of $-$ as the last row. See Table 1 for the 12-run Plackett-Burman design.

It is well known that if a Hadamard matrix exists, then its order n has to be a multiple of 4. A Hadamard matrix H of order n is an $n \times n$ orthogonal matrix with the elements ± 1 whose columns (and rows) are orthogonal to each other. That means $H^T H = nE$ where E is the identity matrix. One can always normalize a Hadamard matrix by sign changes within complete rows so that its first column consists of all 1's. Removing the first column, one obtains a saturated two-level design with n runs and $(n - 1)$ columns, which is a non-regular design and called a Hadamard design. Plackett-Burman designs are special cases of Hadamard designs. Non-regular designs are useful for factor screening and they fill the gaps between regular designs in terms of various run sizes. Unlike regular two-level fractional factorial designs in which any two effects are either orthogonal or fully aliased, non-regular designs exhibit some complex aliasing structure. In a non-regular design, there exist two effects that are partially aliased, meaning that they are neither orthogonal nor fully aliased.

For example, in a non-regular two-level factorial design, a main effect is partially confounded with all the two-factor interactions not involving itself. Because of this complex aliasing structure, non-regular factorial designs were traditionally not advocated when some interactions are potentially important. However Hamada and Wu (1992) showed that some interactions could be detected using non-regular

factorial designs. Hence the arising question is how to select non-regular fractional factorial designs when some interactions are potentially important and need to be estimated. In this article, a new method was proposed and studied to solve this problem.

Method for selecting optimal non-regular factorial designs

Suppose the interest is in estimating all the m main effects and some important two-factor interactions by using a non-regular two-level fractional factorial design. Then the fitted model should include all the m main effects and important two-factor interactions. The fitted model is given by

$$Y = \beta_0 I + X_1 \beta_1 + \varepsilon \quad (1)$$

where Y denotes the vector of n observations, β_0 is the grand mean and I the all $+1$ column, β_1 is the vector of parameters containing all the main effects and important two-factor interactions, X_1 is the corresponding design matrix, and ε is the vector of uncorrelated random errors, assumed to have mean 0 and a constant variance. Because other interactions may not be negligible, the true model can be written as

$$Y = \beta_0 I + X_1 \beta_1 + X_2 \beta_2 + X_3 \beta_3 + \dots + X_m \beta_m + \varepsilon \quad (2)$$

where β_2 is the vector of parameters containing the remaining two-factor interactions and X_2 is the corresponding design matrix, β_k is the vector of parameters containing k -factors interactions and X_k is the corresponding design matrix. The least square estimator $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T Y$ from the fitted model in (1) has expectation (under the true model in (2)),

$$E(\hat{\beta}_1) = \beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T X_3 \beta_3 + \dots + (X_1^T X_1)^{-1} X_1^T X_m \beta_m \quad (3)$$

So the bias of $\hat{\beta}_1$ for estimating β_1 is given by

SELECTION OF NON-REGULAR FRACTIONAL FACTORIAL DESIGNS

Table 1. The 12-run Plackett-Burman design

Run	1	2	3	4	5	6	7	8	9	10	11	Response
1	+1	+1	-1	+1	+1	+1	-1	-1	-1	+1	-1	y_1
2	-1	+1	+1	-1	+1	+1	+1	-1	-1	-1	+1	y_2
3	+1	-1	+1	+1	-1	+1	+1	+1	-1	-1	-1	y_3
4	-1	+1	-1	+1	+1	-1	+1	+1	+1	-1	-1	y_4
5	-1	-1	+1	-1	+1	+1	-1	+1	+1	+1	-1	y_5
6	-1	-1	-1	+1	-1	+1	+1	-1	+1	+1	+1	y_6
7	+1	-1	-1	-1	+1	-1	+1	+1	-1	+1	+1	y_7
8	+1	+1	-1	-1	-1	+1	-1	+1	+1	-1	+1	y_8
9	+1	+1	+1	-1	-1	-1	+1	-1	+1	+1	-1	y_9
10	-1	+1	+1	+1	-1	-1	-1	+1	-1	+1	+1	y_{10}
11	+1	-1	+1	+1	+1	-1	-1	-1	+1	-1	+1	y_{11}
12	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	y_{12}

$$\text{Bias}(\hat{\beta}_1, \beta_1) = (X_1^T X_1)^{-1} X_1^T X_2 \beta_2 + (X_1^T X_1)^{-1} X_1^T X_3 \beta_3 + \dots + (X_1^T X_1)^{-1} X_1^T X_m \beta_m \quad (4)$$

The matrix $A_2 = (X_1^T X_1)^{-1} X_1^T X_2$ is called the second alias matrix, $A_k = (X_1^T X_1)^{-1} X_1^T X_k$ the k th alias matrix. The idea of alias matrix was

originally introduced by Box and Wilson (1951). In other words, A_2 shows to which extent the estimates of the main effects and the important two-factor interactions in the model will be biased by the remaining two-factor interactions, A_k shows to which extent the estimates of the effects in the model will be biased by the k -factor interactions.

Note that $A_2 \beta_2$ is the contribution of β_2 to the bias and $A_k \beta_k$ is the contribution of β_k to the bias. Because β_k is unknown, we have to work with A_k and minimize $A_k \beta_k$ through minimizing A_k . One size measure for a matrix A

$$= (a_{ij}) \text{ is given by } \|A\|^2 \stackrel{\text{def}}{=} \text{trace}(A^T A) = \sum_{i,j} a_{ij}^2.$$

Under the hierarchical assumption that lower order effects are more important than higher order effects, to minimize the bias of $\hat{\beta}_1$ we should sequentially minimize $\|A_2\|^2, \dots, \|A_m\|^2$. Here $\|A_k\|^2$ can be viewed as a confounding index which is a measure of the partial confounding between j -factor interactions not in

the model and the effects in the model. For regular two-level fractional factorial designs, the entries of A_k are 0 or 1, and thus $\|A_k\|^2$ is simply the number of k -factor interactions not in the model confounded with the effects in the postulated model (Ke & Tang, 2003). For non-regular two-level fractional factorial designs, the entries of A_k are usually not integers because of the partial confounding structure. Now let $N_k = \|A_k\|^2$. Based on the above results, we can select optimal non-regular two-level fractional factorial designs by sequentially minimizing $N_2, \dots, N_m$ where N_k is a measure of the bias contributed by the k -factor interactions. The design selection criterion is given below.

Optimal design selection criterion: Suppose the interest is in estimating all the m main effects and some important two-factor interactions by using a non-regular two-level fractional factorial design. Let $A_k, k = 2, 3, \dots, m$ be the k th alias matrix of the model and let $N_k = \text{trace}(A_k^T A_k)$ which is a measure of A_k . The optimal design is selected by sequentially minimizing $N_2, \dots, N_m$.

To gain further insight into the criterion, examine the criterion in detail. The postulated model consists of all the main effects and important two-factor interactions. If the effects not in the postulated model cannot be

Figure 1. Graph for model with one 2-factor interaction.

Figure 2. Graphs for models with two 2-factor interactions.

Figure 3. Graphs for models with three 2-factor interactions.

completely ignored, they will bias the estimates of the effects in the model. To solve this problem, the key issues are to permit estimation of the main effects and important two-factor interactions in the postulated model and to minimize the bias caused by the effects not in the model. Those two-factor interactions not in the model and higher-order interactions generally cause a bias on the estimation of the effects in the model. The measure of this bias, as given by N_k , is a measure of the bias caused by the k -factor interactions. Under the hierarchical principle that lower-order effects are more important than higher-order effects (Wu & Hamada, 2000), to minimize the bias, we should sequentially minimize $N_2, N_3, \dots, N_m$. The vector $(N_2, N_3, \dots, N_m)$ is called the confounding index pattern of a design. The optimal design should be selected such that to sequentially minimize the bias caused by those non-negligible interactions. Therefore this criterion selects the optimal non-regular design that has minimum N_2 . If several designs have the same number of N_2 , it selects optimal design that has minimum N_3 among the designs that have minimum N_2 , and so on.

Results

Searching method

Consider 12-run Plackett-Burman design as an example. Let k be the number of important two-factor interactions. For $k = 1$, there is only one non-isomorphic model, as represented by Figure 1. For $k = 2$ and 3, the number of non-isomorphic models is 2 and 5 and the graphs for these non-isomorphic models are given in Figure 2 and 3 respectively.

Because there are many choices for the assignment of the important two-factor interactions, the optimal Plackett-Burman design of 12 runs is not easy to select according to this criterion. A computer program is used to calculate the confounding index pattern for each choice of the designs for each model for the given number of main effects and important two-factor interactions. Then select the best one that has minimum N_2 . If several designs have same N_2 , we select the best one that has the minimum N_3 , and so on.

For 12-run Plackett-Burman designs, Draper (1985) and Wang (1989) showed that except for $m = 5$ and 6, any $12 \times m$ designs are equivalent. Lin and Draper (1992) and Wang and Wu (1995) showed that the two non-

SELECTION OF NON-REGULAR FRACTIONAL FACTORIAL DESIGNS

Table 2. Optimal 12-run Plackett-Burman designs for the model containing one two-factor interaction

m	design columns	2-f interaction	(N_2, N_3, N_4)
4	1 2 3 4	(1, 2)	(1.56, 1.01, 0.67)
5	1 2 3 4 5	(1, 2)	(4.48, 4.30, 3.26)
6	1 2 3 4 5 6	(1, 2)	(9.06, 13.16, 11.91)
7	1 2 3 4 5 6 7	(1, 6)	(18.22, 29.56, 30.44)
8	1 2 3 4 5 6 7 8	(1, 7)	(32.78, 62.67, 72)
9	1 2 3 4 5 6 7 8 9	(1, 2)	(59, 130.67, 160)
10	1 2 3 4 5 6 7 8 9 10	(1, 2)	(124, 320, 400)

Table 3. Optimal 12-run Plackett-Burman designs for the models containing two two-factor interactions

m	model	design columns	2-f interactions	(N_2, N_3, N_4)
4	2(a)	1 2 3 4	(1, 2)(3, 4)	(1.16, 3.34, 0.72)
	2(b)	1 2 3 4	(1, 2)(1, 3)	(1.65, 1.72, 1.04)
5	2(a)	1 2 3 4 5	(1, 3)(2, 4)	(5.46, 7.38, 4.54)
	2(b)	1 2 3 4 5	(1, 2)(2, 3)	(5.25, 8.08, 5.08)
6	2(a)	1 2 3 4 5 6	(1, 4)(2, 3)	(12.56, 18.78, 14.44)
	2(b)	1 2 3 4 5 6	(1, 2)(2, 4)	(12.22, 20.44, 15.33)

Table 4. Optimal 12-run Plackett-Burman designs for the models containing three two-factor interactions

m	model	design columns	2-f interactions	(N_2, N_3, N_4)
4	3(c)	1 2 3 4	(1, 2)(1, 3)(1, 4)	(1.41, 2.75, 1.67)
	3(d)	1 2 3 4	(1, 4)(4, 3)(3, 2)	(1.24, 4.05, 1.13)
	3(e)	1 2 3 4	(1, 2)(1, 3)(2, 3)	(1.41, 2.75, 1.67)
5	3(b)	1 2 3 4 5	(1, 2)(3, 5)(4, 5)	(6.33, 11.29, 6.33)
	3(c)	1 2 3 4 5	(2, 3)(2, 4)(2, 5)	(6.40, 10.16, 6.12)
	3(d)	1 2 3 4 5	(3, 2)(2, 1)(1, 4)	(6.20, 11.82, 6.76)
	3(e)	1 2 3 4 5	(1, 2)(1, 3)(2, 3)	(6.40, 10.16, 6.12)
6	3(a)	1 2 3 4 5 6	(1, 2)(3, 5)(4, 6)	(19.78, 33.33, 22.22)
	3(b)	1 2 3 4 5 6	(3, 6)(1, 4)(2, 4)	(16.39, 25.05, 17.17)
	3(c)	1 2 3 4 5 6	(1, 2)(2, 3)(2, 4)	(15.76, 28.19, 18.98)
	3(d)	1 2 3 4 5 6	(1, 4)(4, 2)(2, 3)	(15.82, 26.48, 18.13)
	3(e)	1 2 3 4 5 6	(1, 2)(1, 3)(2, 3)	(15.95, 27.05, 18.41)

isomorphic 12×5 designs are the sub-matrix of the saturated design in Table 1 consisting of columns 1, 2, 3, 4, and 5 and the one consisting of columns 1, 2, 3, 4, and 10, the two non-isomorphic 12×6 designs are the sub-matrix consisting of columns 1, 2, 3, 4, 5 and 6 and the one consisting of columns 1, 2, 3, 4, 5 and 7. The estimation capacity of the 12-run Plackett-Burman designs were studied by Li and Wang (2004). They proved that the design with $m = 4$ can estimate all two-factor interactions, the designs with $7 \leq m \leq 10$ can only estimate any one two-factor interactions, and the two designs of 12×5 and one design of 12×6 (with columns 1, 2, 3, 4, 5 and 6) can estimate any models with up to three two-factor interactions. The above information can help us to save time and effort for searching the optimal 12-run Plackett-Burman designs. We have found the optimal 12-run Plackett-Burman designs for the models containing up to three two-factor interactions using (N_2, N_3, N_4) instead of the entire vector $(N_2, \dots, N_m)$ to further reduce the computing burden. Actually five-factor and higher order interactions are very small and usually negligible in practice.

Optimal 12-run Plackett-Burman designs

Tables 2, 3, and 4 present optimal 12-run Plackett-Burman designs for the models with one, two, and three important two-factor interactions respectively. In these tables, the entries under “ m ” give the number of factors, the entries under “model” indicate which model is under consideration, and for example, an entry of 2(a) denotes the model represented by Figure 2(a). The entries under “design columns” give the design columns of for the factors in the fitted model. Column j in these tables denotes the j -th column in the saturated 12-run Plackett-Burman design in Table 1. The entries under “2-f interaction” show how to assign the factors involved in the important two-factor interactions. The last column in these tables gives (N_2, N_3, N_4) .

The optimal 12-run Plackett-Burman designs are listed in Tables 2–4. When planning to study several factors and some important two-factor interactions by using a 12-run Plackett-Burman design, choose an optimal design directly from these tables to satisfy the current

needs. Now an example is employed to illustrate how to use these optimal design tables.

Suppose that in an experiment, the experimenter want to study six factors, temperature, moisture, light, nitrogen, phosphorus, and potassium. Suppose a 12-run Plackett-Burman design is being considered. In addition to the main effects of these factors, suppose further there is the need to estimate the three two-factor interactions that are between temperature and nitrogen, between temperature and phosphorus, and between temperature and potassium. The graph for this model is 3(c) as in Figure 3. The optimal 12-run Plackett-Burman design for this model can be found at the row for $m = 6$ and model 3(c) in Table 4. From this row in Table 4, we see that the design columns are 1, 2, 3, 4, 5, and 6 in Table 1. To complete the specification of the optimal design, the six factors need to be appropriately assigned to the six columns. The 2-f interaction column in Table 4 says that we should assign temperature to column 2, and assign nitrogen, phosphorus, and potassium to column 1, 3, and 4 arbitrarily. Other two factors can be arbitrarily assigned to the remaining columns 5 and 6. This design has $N_2 = 15.76$ which is a measure of the bias caused by the two-factor interactions, meaning that this design is the best in the sense that no other designs have smaller N_2 than this one for the given model.

References

- Box, G. E. P. & Bisgaard, S. (1993). George’s column. What can you find out from twelve experimental runs? *Quality Engineering*, 5, 663-668.
- Box, G. E. P. & Tyssedal, J. (1996). Projective properties of certain orthogonal arrays. *Biometrika*, 83, 950-955.
- Box, G. E .P. & Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the royal statistical society, Series B*, 13, 1-45.
- Butler, N. A. (2003). Minimum aberration construction results for non-regular two-level fractional factorial designs. *Biometrika*, 90, 4, pp. 891-898.

SELECTION OF NON-REGULAR FRACTIONAL FACTORIAL DESIGNS

- Cheng, C. S. (1995). Some projection properties of orthogonal arrays, *The Annals of Statistics*, 23, 1223–1233.
- Cheng, C. S. (1998). Some hidden projection properties of orthogonal arrays with strength three. *Biometrika*, 85, 491–495
- Cheng, S. W., Li, W., & Ye, K. Q. (2004). Blocked non-regular two-level factorial designs, *Technometrics*, 46, 269–279.
- Dean, A. M. & Draper, N. R. (1999). Saturated main-effect designs for factorial experiments. *Statistics and Computing*, 9, 179–185.
- Deng, L.-Y., Li, Y., & Tang, B. (2000). Catalogue of small run non-regular designs from Hadamard matrices with generalized minimum aberration. *Communications and Statistics, A*, 29, 1379–95.
- Deng, L.-Y. & Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Statistica Sinica*, 9, 1071–1082.
- Deng, L.-Y. & Tang, B. (2002). Design selection and classification for Hadamard matrices using generalized minimum aberration criteria. *Technometrics*, 44, 173–184.
- Draper, N. R. (1985). Small composite designs. *Technometrics*, 27, 173–180.
- Hamada, M. & Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24, 130–137.
- Ke, W. & Tang, B. (2003). Selecting 2^{m-p} designs using a minimum aberration criterion when some two-factor interactions are important. *Technometrics, Volume 45, Number 4, November 2003*, pp. 352–360.
- Li, Y. & Wang, J. (2004). Non-regular designs from Hadamard matrices and their estimation capacity. *Metrika*, 60, 295–303.
- Lin, D. K. J. & Draper N. R. (1992). Projection properties of Plackett and Burman designs. *Technometrics*, 34, 423–428.
- Lin, D. K. J. & Draper N. R. (1993). Generating alias relationships for two-level Plackett and Burman designs. *Computational Statistics and data analysis*, 15, 147–157.
- Lin, D. K. J. & Draper N. R. (1995). Screening properties of certain two-level designs. *Metrika*, 42, 99–118
- Ma, C.-X. & Fang, K.-T. (2001). A note on generalized aberration in factorial designs. *Metrika*, 53, 85–93.
- Plackett, R. L. & Burman, J. P. (1946). The Design of optimum multi-factorial experiments. *Biometrika*, 33, 305–325.
- Tang, B. (2001). Theory of J-characteristics for fractional factorial designs and projection justification of minimum G_2 -aberration. *Biometrika*, 88, 401–7.
- Tang, B. & Deng, L.-Y. (1999). Minimum G_2 -aberration for non-regular fractional factorial designs. *Annals of Statistics*, 27, 1914–1926.
- Wang, J. C. (1989). Orthogonal arrays and nearly orthogonal arrays with mixed levels: Construction and applications. *Unpublished Ph.D: thesis. Department of Statistics, University of Wisconsin-Madison.*
- Wang, J. C. & Wu, C. F. J. (1995) A hidden projection property of Plackett-Burman and related designs. *Statistica Sinica*, 5, 235–250.
- Wu, C. F. J. & Hamada, M. (2000). *Experiments: Planning, Analysis, Parameter Design Optimization*. New York: Wiley.
- Xu, H. & Wu, C. F. J. (2001). Generalized minimum aberration for asymmetric fractional factorial designs. *Annals of Statistics*, 29, 1066–1077.

A Monte Carlo Power Analysis of Traditional Repeated Measures and Hierarchical Multivariate Linear Models in Longitudinal Data Analysis

Hua Fang
University of Nebraska-Lincoln

Gordon P. Brooks
Ohio University

Maria L. Rizzo
Bowling Green State University

Kimberly A. Espy
University of Nebraska-Lincoln

Robert S. Barcikowski
Ohio University

The power properties of traditional repeated measures and hierarchical linear models have not been clearly determined in the balanced design for longitudinal studies in the current literature. A Monte Carlo power analysis of traditional repeated measures and hierarchical multivariate linear models are presented under three variance-covariance structures. Results suggest that traditional repeated measures have higher power than hierarchical linear models for main effects, but lower power for interaction effects. Significant power differences are also exhibited when power is compared across different covariance structures. Results also supplement more comprehensive empirical indexes for estimating model precision via bootstrap estimates and the approximate power for both main effects and interaction tests under standard model assumptions.

Key Words: Monte Carlo, power analysis, traditional repeated measures, hierarchical multivariate linear models, longitudinal study.

Introduction

In longitudinal studies, both traditional repeated measures (TRM) and hierarchical multivariate linear models (HMLM) can be applied for a balanced design when the focus is testing fixed main effects. The balanced design assumes an equal number and spacing of measurements over time for each subject. TRM can be used for this

design with univariate or multivariate approaches. When sphericity is met, the univariate tests are appropriate; when sphericity is not met, we can employ adjusted univariate tests or traditional multivariate tests, which do not assume the variance-covariance (VC) structure (cf., Greenhouse & Geisser, 1959; Huynh & Feldt, 1976; Jennrich & Schluchter, 1986; Wolfinger & Chang, 1995). For the same longitudinal design, HMLM treat the repeated observations nested within the subjects, that is, repeated measures at level-1 and subjects at level-2. A third or higher level of HMLM can be introduced to represent the contextual effects on the subjects' growth (Raudenbush & Bryk, 2002).

Hua Fang is a Research Assistant Professor and Project Director in the Office of Research. E-mail: jfang2@unlnotes.unl.edu. Gordon P. Brooks is an Associate Professor of Research and Evaluation. E-mail: brooksg@ohio.edu. Maria L. Rizzo is an Assistant Professor. Email her at mrizzo@bgnet.bgsu.edu. Kimberly Andrews Espy is a Professor in the Department of Psychology. Robert S. Barcikowski is an emeritus professor of Research and Evaluation. E-mail: barcikow@ohio.edu.

HMLM and TRM are essentially interrelated in their theoretical development, especially after advanced computational methods were developed to handle missing values and model the VC structures (Dempster, Laird & Rubin, 1977; Dempster, Rubin & Tsutakawa, 1981; Goldstein 1995; Jennrich & Schluchter, 1986; Littell, Milliken, Stroup &

Wolfinger, 2006; Little, 1995; Little & Rubin, 2002; Maas & Snijders, 2003; McCulloch & Searle, 2001; Raudenbush & Bryk, 2002; Van der Leenden, Vrijburg & de Leeuw, 1996). Jennrich and Schluchter were the first to model specific VC structures directly through maximum likelihood estimation based on traditional multivariate repeated measures approach whereas HMLM incorporates Jennrich and Schluchter's multivariate repeated measures approach to longitudinal data analysis (Schluchter, 1988; Van der Leenden, 1998; Jennrich & Schluchter, 1986; Raudenbush & Bryk, 2002). In literature, HMLM is simply called hierarchical linear models, more generally known as multilevel models, growth mixture models or generalized latent variable models (e.g., Goldstein, 1994, 1995; Hox, 2002; Maas & Snijders, 2003; Muthén, 2002, 2004; Muthén & Muthén, 2006; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Skrondal & Rabe-Hesketh, 2004).

The power analysis in longitudinal studies has been an active area but a uniform and standard criterion has not been established, especially based on the VC structures (cf., Hedeker, Gibbons & Waternaux, 1999; Littell et al., 2006; Raudenbush et al., 2005; Snijders, 2005). As the current analytical power approximations are not comprehensive or necessarily accurate (Littell et al., 2006) and the power properties of TRM and HMLM have not yet been clearly compared in the balanced design, using Monte Carlo (MC) simulation approach would be efficient to examine their power properties simultaneously.

For parsimonious and exploratory purposes, TRM and three common VC structures were examined with the longitudinal data generated from a 2-level HMLM in this study. The three VC structures were: (a) Random slope with homogeneous level-1 variance (RC); (b) unstructured (UN); (c) and first-order autoregressive (AR(1)). Additionally, the bootstrap estimates for the treatment effect were compared for TRM and the three VC structures.

Two-level HMLM model

The hypotheses tested in this simulation assumed no fixed effects on the individuals'

scores over time. The fixed effects were the two-group treatment effect (β_{01}), time effect (β_{10}) and interaction (β_{11}). The underlying mathematical model for this simulation is as follows:

$$\text{Level 1: } y_{it} = \pi_{0i} + \pi_{1i} * \text{TIME} + e_{it} \quad (1)$$

$$\begin{aligned} \text{Level 2: } \pi_{0i} &= \beta_{00} + \beta_{01} * \text{TREATMENT} + u_{0i} \\ \pi_{1i} &= \beta_{10} + \beta_{11} * \text{TREATMENT} + u_{1i} \end{aligned} \quad (2)$$

where y_{it} represents the score of person i at time t ; π_{0i} is the score of person i at time 0; π_{1i} refers to the slope of person i (i.e., rate of change with respect to time); β_{00} is the average overall initial score at time 0; β_{01} stands for the hypothesized difference in average status from the effect of treatment; β_{10} is the average overall annual rate of change at level-2; β_{11} represents the hypothesized difference in average annual rate of change from the effect of treatment; u_{0i} is the random effect for intercepts (i.e., random error of intercepts at level-2); u_{1i} is the random effect for slopes (i.e., random error of slopes at level-2); e_{it} refers to the random error at the t^{th} time point of the i^{th} person at level-1.

The above 2-level model can be reduced to a single level model by substituting Equation (2) into (1):

$$y_{it} = (\beta_{00} + \beta_{01} \times \text{TREATMENT} + \beta_{10} \times \text{TIME} + \beta_{11} \times \text{TREATMENT} \times \text{TIME}) + r_{it} \quad (3)$$

where the residual term, $r_{it} = u_{0i} + u_{1i} * \text{TIME} + e_{it}$, includes the level-1 random error (e_{it}) and level-2 random effects (u_{0i} and u_{1i}); β_{00} , β_{01} , β_{10} , β_{11} , u_{0i} , u_{1i} , and e_{it} are the same as those in Equation (1) and Equation (2). Hence, the HMLM model is also expressed as the mixed effect model with a mix of fixed effects in the parenthesis and random effects embodied in the residual term r_{it} .

TRM and three covariance structures under study

The TRM approach to equation (3) can be simply expressed in a matrix form:

$$Y = X\beta + r \quad (4)$$

where Y is a $t_i \times 1$ response vector for subject i , and t represents the number of time points and $i = 1, \dots, n$; X is a $t_i \times a$ design matrix for fixed effect β , where a is the number of fixed effects (i.e., the 3 parameters, β_{01} , β_{10} , and β_{11} in this study), and β is an $a \times 1$ vector; residual r is independently and normally distributed with a mean vector of 0 and variance of Σ , $r \sim N(0, \Sigma)$. The parameter estimates in the traditional approach are obtained using the method of moments (McCulloch & Searle, 2001; Montgomery, 2005; Wolfinger & Chang, 1995).

Random slope with homogeneous level-1 variance (RC)

Random slope with homogeneous level-1 variance is often described as the covariance structure for standard MLM, also known as standard hierarchical linear model (HLM) or random coefficient model (RC) (Raudenbush & Bryk, 2002, p. 191; Raudenbush, Bryk & Congdon, 2004; Singer & Willett, 2003, p.244-245, 251-265; Kreft, 1996). For convenience, RC is used for this covariance structure hereafter. The RC covariance structure of Model (3) residual r_{ti} , Σ_r , is expressed as two components:

$$e_{ti} \sim N(0, \sigma^2), \text{ and} \\ \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \right].$$

The variance of level-1 error term (e_{ti}) is homogeneous and the covariance structure of level-2 random errors (u_{0i} and u_{1i}) is arbitrary. For Model (3), only 4 variance-covariance parameters need to be estimated, that is, σ^2 , τ_{00} , τ_{11} and τ_{01} . Level-1 variance, σ^2 , is independent of level-2 variance, τ .

Unstructured Covariance Matrix (UN)

The unstructured covariance matrix (also called unrestricted structure in literature) places no restrictions on the structure of covariance matrix, Σ_r , and there is redundancy in mathematical formulation of this covariance structure (Littell, Henry, & Ammerman, 1998, pp. 1229-1230; Raudenbush et al., 2004). If the

covariance structure of Σ_r is assumed unknown, one could fit an UN covariance matrix. The UN matrix for each level-2 subject with 3 time points can be expressed as

$$\begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 \end{pmatrix} \quad (5)$$

and requires the estimation of 3 variance parameters and 3 covariance parameters. When more time points are involved, UN can require an exorbitant number of parameters.

First Order Auto-Regressive (AR(1))

For Model (3), AR(1) can be written as follows:

$$\begin{aligned} Var(r_{ti}) &= \tau + \sigma^2 \\ Cov(r_{ti}, r_{t'i'}) &= \tau + \sigma^2 \rho^{|t-t'|} \end{aligned} \quad (6)$$

where τ stands for the level-2 variance and $|t - t'|$ is the lag between two time points; ρ is the auto-correlation and σ^2 is the level-1 variance at each time point. AR(1) allows the level-1 errors to be correlated under Markov assumptions and level-1 covariance structure is expressed as

$$\sigma^2 \begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix} \quad (7)$$

As the redundancy is not in the mathematical formulation of AR(1), the covariance structure of level-2 random effects (u_{0i} and u_{1i}) must be specified to estimate the level-2 variance τ (τ_{00} , τ_{11} and τ_{01}) which is usually assumed unstructured (cf. Littell, Henry, & Ammerman, 1998, pp. 1229-1230; McCulloch & Searle, 2001; Raudenbush & Bryk, 2002; Singer & Willett, 2003; Wolfinger, 1993). Thus, 5 variance-covariance parameters need to be estimated for AR(1) of Model (3).

Methodology

Monte Carlo Design

This research employed a Monte Carlo (MC) study to compare the empirical power of RC, UN and AR(1). To make the results applicable over many possible situations, a standardized model, Model (4), was employed in this simulation where the grand mean in Model (3) was set to zero (i.e., $\beta_{00} = 0$)

$$y_{ti} = (\beta_{01} \times \text{TREATMENT} + \beta_{10} \times \text{TIME} + \beta_{11} \times \text{TREATMENT} \times \text{TIME}) + r_{ti} \quad (8)$$

A stacked SAS macro was written by the author (the author, 2006) to generate the two-level repeated measures data with the RC covariance structure and to calculate the power for RC, UN and AR(1). The number of iterations for this MC study was 5000, and the nominal alpha (α) for each sample test was .10 considering the relatively small number of iterations (e.g., compared to 10,000 iterations).

Data generation

The data generation procedure based on Model (3) was carried out as follows:

Level-1 data

The error term at level-1 (i.e., e_{ti}) was assumed to be independent of the level-2 random effects (i.e., u_{0i} and u_{1i}), that is, $\text{cov}(u_i, e_{ti}) = 0$. The level-1 error term followed a normal distribution, $e_{ti} \sim N(0, \sigma^2)$.

Level-2 data

The random intercepts u_{0i} ($X_{intercepts}$), and slopes u_{1i} (X_{slopes}), assumed a standard bivariate normal distribution. A standardized G matrix for

$$X_{intercepts} \text{ and } X_{slopes}, G = \begin{pmatrix} \sigma_{00}^2 & \sigma_{01}^2 \\ \sigma_{10}^2 & \sigma_{11}^2 \end{pmatrix}, \text{ and random}$$

mean vector, $\begin{pmatrix} \mu_0 \\ \mu_1 \end{pmatrix}$, were specified to simulate

correlated bivariate normal data for $X_{intercepts}$ and X_{slopes} . The Cholesky decomposition method was utilized to generate the correlated level-2 normal data. This simulation was accomplished by multiplying the normal data by L which is the

Cholesky decomposition of G . The estimated variables were $\hat{X}_{intercepts}$ and $\hat{X}_{slopes}$.

Complete data

Data were generated in the appropriate format required by PROC MIXED (SAS Institute Inc., 2003). An index matrix was created for time, treatment and individual IDs. Two treatment groups were coded by 0 and 1, respectively. Individuals (IDs) were considered nested within each treatment group, for instance, IDs ranged from 1 to 25 for Group 1, and 26 to 50 for Group 2. Time started from 0 and extended to the maximum specified for each study condition. Based on Model (4), a univariate response vector of y_{ti} was created. For example, each subject might have had 3 time points and each treatment group had 25 subjects.

The data generator (author, 2006) was validated with parameter estimates from Potthoff and Roy's data (1964). The results are shown in Appendix A.

Power comparison

Holding other factors constant, the power comparison was implemented by changing the levels of one of the four factors, respectively: (1) Correlation in G matrix (G), (2) reliability of level-1 coefficients (λ), (3) effect size (β) and (4) ratio of group sample size to time points (n/t) under specified conditions (see Table 1).

Power comparison by G matrix

This study used Cohen's indices (1988) for correlation, $\rho \in \{.1 \ .3 \ .5\}$; correspondingly,

G matrix (G) for random intercepts (u_{0i}) and slopes (u_{1i}) was specified as $\begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$,

or $\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$. To show the power pattern by

varying correlation in G matrix, the MC design incorporated a moderate sample size ($n = 75$) and fixed time points ($t = 3$) (i.e., ratio of group sample size to time points, $n/t = 75/3$), effect size ($\beta_{01} = .5$, $\beta_{10} = .5$ or $\beta_{11} = .5$), and moderate reliability $\lambda = .5$, to simulate a specific situation and compare power at each G matrix. Based on the design, a general power pattern of TRM, RC,

UN and AR(1) were presented at each of the three G matrices in the respective three tests, treatment effect (β_{01}) test, time effect (β_{10}) test and interaction (β_{11}) test (i.e., $4 \times 3 \times 3 = 36$ cells).

Power comparison by reliability

By changing the averaged reliability of level-1 coefficients, $\lambda \in \{.01, .25, .5, .75, .1\}$, (Raudenbush & Bryk, 2002), the power pattern of RC, UN and AR(1) was compared by setting $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$, $n/t = 75/3$, and $\beta = .5$. The power pattern of TRM, RC, UN and AR(1) was presented at each of the five reliability indexes in the respective three tests, treatment effect (β_{01}), time effect (β_{10}) and interaction (β_{11}) (i.e., $4 \times 5 \times 3 = 60$ cells).

Power comparison by effect size

Cohen's indexes were also used for two-group treatment effect (β_{01}), time effect (β_{10}) and interaction (β_{11}), $\beta \in \{.2, .5, .8\}$. The MC design simulated a moderate situation where $n = 75$, $C_x = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$, $\lambda = .5$ and $t = 3$ to compare power at three effect sizes, $\beta \in \{.2, .5, .8\}$, of the three fixed effects for the four models, TRM, RC, AR(1) and UN (i.e., $4 \times 3 \times 3 = 36$ cells).

Power comparison by sample size ratio

For exploratory purpose, this study fixed the time points (t) at 3. As the maximum likelihood estimation requires relatively large sample sizes, the sample size per treatment group ($m=2$) was changed from 25 to 200 by an increase of 25 ($n \in \{25, 75, 100, 125, 150, 175, 200\}$), that is, the total sample size $N \in \{150, 300, 450, 600, 750, 900, 1050, 1200\}$ ($N = m \times n \times t$). To compare the power by varying the sample size ratio, the condition was specified as $\beta = .5$, $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$, $\lambda = 3$ and $t = 3$. For each specified condition, the power patterns for TRM and the three VC structures were presented at eight sample sizes, for the three fixed effects (i.e., $4 \times 8 \times 3 = 96$ cells).

Monte Carlo Analysis

The following function was employed to calculate the upper bound of standard errors for pairwise empirical power (i.e., the standard error for the difference in proportions).

$$SE_{upper} = \sqrt{2 \times \frac{p \times (1-p)}{n}} \quad (9)$$

where $p = .5$ and $n = 5000$. If the pairwise differences are twice the upper bound of SE (i.e., $SE_{upper} = .20$), then the differences are labeled as significant. The power patterns are illustrated in tables and graphs. In addition to the power analysis, bootstrap standard CI, estimates, bias and standard errors for the estimates of the treatment mean difference (β_{01}) were calculated to compare the model precision.

Results

Empirical Power by G Matrix

The results (see Figure 1) indicated two general patterns when varying the *G matrix* under the specific circumstance in all three tests, treatment effect (β_{01}), time effect (β_{10}) and interaction (β_{11}). The first pattern showed that as the correlation in *G matrix* increases, the power of TRM, RC, UN and AR(1) decreased slightly, which may imply that the lower the correlation between intercepts and slopes, the higher the power we can obtain. But it should be noted that the power change across G matrices seems to be minimal.

The pairwise power tests (see Table 2) showed that TRM power was significantly higher than the other three in the treatment and time tests, but significantly lower than the other three in the interaction test by varying G matrices. Among the three VC structures, UN had significantly higher power than RC and AR(1) in both treatment and interaction tests, whereas AR(1) has significantly higher power than RC in the same two tests. As to the time test, UN power was significantly higher than RC across the three G matrices but was not significantly higher than AR(1) at all three G matrices.

REPEATED MEASURES AND HMLM IN LONGITUDINAL DATA ANALYSIS

Table 1. MC Design for Power Analysis of TRM, RC, AR(1) and UN by G Matrix, Effect Size and Sample Size of 5000 MC Samples at $\alpha = .10$

Factors (Cells) ^a	Conditions ^b
<p>G Matrix (G) (36)</p> $\left\{ \begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix} \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix} \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} \right\}$	<p>Fixed^c: $n = 75, t = 3$ and $\lambda = 3$</p> <ol style="list-style-type: none"> 1. $\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$^d 2. $\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$^d 3. $\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = .5$^d
<p>Reliability (λ) (60)</p> $\{.01 \ .25 \ .5 \ .75 \ 1\}$	<p>Fixed^c: $t = 3, \lambda = 3$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$</p> <ol style="list-style-type: none"> 1. $\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$^d 2. $\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$^d 3. $\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = .5$^d
<p>Effect Size (β) (36)</p> $\{.2 \ .5 \ .8\}$	<p>Fixed^c: $n = 75, t = 3, \lambda = 3$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$</p> <ol style="list-style-type: none"> 1. $\beta_{10} = 0, \beta_{11} = 0$^d 2. $\beta_{01} = 0, \beta_{11} = 0$^d 3. $\beta_{01} = 0, \beta_{10} = 0$^d
<p>Sample Size per Treatment Group (n) (96)</p> $\{25 \ 50 \ 75 \ 100 \ 125 \ 150 \ 175 \ 200\}$	<p>Fixed^c: $t = 3, \lambda = 3$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$</p> <ol style="list-style-type: none"> 1. $\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$^d 2. $\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$^d 3. $\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = .5$^d

Note. ^a The factors are not crossed

^b Conditions are specified for testing each fixed effect (β_{01} , β_{10} , and β_{11}) within each factor

^c "Fixed" indicates the fixed parameters in the design within each factor

^d Settings for testing the three fixed effects

Empirical Power by Reliability

Two power patterns were displayed by varying reliability indexes in the specified condition. First, with the increase of reliability, the power of TRM, RC, AR(1) and UN increased in all three tests. TRM had a higher power-increasing rate than the other three below the reliability of .5. Above the reliability of .5, all four seemed to increase power at a decreasing rate. The power of all four approached to the asymptote of 1 as the reliability reached 1 (see Figure 2)..80. In the time test, all four had power above .80. Yet, in the interaction test, TRM power was significant in all three tests, that is, TRM had significantly

higher power than the other three, reliability reached 1 (see Figure 2).

The second pattern showed that TRM gained the highest power in the treatment and time tests but had the lowest power than the other three in the interaction test across the reliability indexes (see Table 3). Among RC, AR(1) and UN, UN power ranked the highest, AR(1) the second and RC the lowest across all reliability indexes in all three tests. At the reliability of .75, TRM power was above .80 in the treatment test whereas the power of all three VC structures seemed to be above .60 but below .80. In the time test, all four had power above .80. Yet, in the interaction test, TRM power was

Figure 1. Power pattern of HLM, AR(1), UN and TRM by G matrix in treatment, time and interaction tests when $n = 75$, $t = 3$ and $\lambda = .5$ of 5000 MC samples at $\alpha = .10$.

below .60 while the power of the other three was at or above .60.

Table 4 indicated that when the reliability was between .5 and 1, the pairwise power differences among the four were significant in all three tests, that is, TRM had significantly higher power than the other three, respectively, in the treatment and time tests, but significantly lower power than the three in the interaction test. Generally, among the three VC structures, UN power was statistically higher than AR(1) and RC while AR(1) power was significantly higher than RC under the specified condition. Below the reliability of .5, the pairwise power differences among the four were not all significant across the three tests.

Empirical Power by Effect Size

As effect sizes increased, the power of the four models was enhanced in the treatment, time and interaction tests. Also, it seemed that the power of the four has a higher increasing rate from the small to the medium effect size than from the medium to the large effect size (Fig. 3).

Table 4 showed the significant pairwise power differences among TRM, RC, AR(1) and UN at the medium effect size in all three tests: TRM power was significantly higher than the other three VC structures in treatment and time tests, but had significantly lower power than the other three in the interaction test. Still at the medium effect size, UN was significantly higher than RC and AR(1) while AR(1) was

REPEATED MEASURES AND HMLM IN LONGITUDINAL DATA ANALYSIS

Table 3. Power and Pairwise Power Difference of TRM, RC, AR(1) and UN by Reliability when $n = 75$, $t = 3$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$ of 5000 MC Samples at $\alpha = .10$

λ	Power					Power Difference				
	RC	AR(1)	UN	TRM	UN vs. RC	UN vs. AR(1)	AR(1) vs. RC	TRM vs. RC	TRM vs. AR(1)	TRM vs. UN
Two-group Treatment Effect ($\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$)										
.01	.1050	.1092	.1166	.1148	.0116	.0074	.0042	.0098	.0056	-.0018
.25	.2436	.2650	.2726	.4472	.0290*	.0076	.0214*	.2036*	.1822*	.1746*
.50	.4264	.4688	.4946	.6762	.0682*	.0258*	.0424*	.2498*	.2074*	.1816*
.75	.6054	.6646	.7292	.8208	.1238*	.0646*	.0592*	.2154*	.1562*	.0916*
1	.7830	.8374	.9334	.8920	.1504*	.0960*	.0544*	.1090*	.0546*	-.0414*
Time Effect ($\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$)										
.01	.1162	.1200	.1216	.1292	.0054	.0016	.0038	.0130	.0092	.0076
.25	.5210	.5338	.5418	.7050	.0208*	.0080	.0128	.1840*	.1712*	.1632*
.50	.7354	.7576	.7828	.9284	.0474*	.0252*	.0222*	.1930*	.1708*	.1456*
.75	.8586	.8776	.9208	.9838	.0622*	.0432*	.0190*	.1252*	.1062*	.0630*
1	.9138	.9330	.9764	.9962	.0626*	.0434*	.0192*	.0824*	.0632*	.0198*
Interaction ($\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = 0.5$)										
.01	.1092	.1146	.1194	.1092	.0102	.0048	.0054	.0000	-.0054	-.0102
.25	.3222	.3336	.3416	.2796	.0194*	.0080	.0114	-.0426*	-.0540*	-.0620*
.50	.4844	.5092	.5390	.4368	.0546*	.0298*	.0248*	-.0476*	-.0724*	-.1022*
.75	.5972	.6348	.6952	.5576	.0980*	.0604*	.0376*	-.0396*	-.0772*	-.1376*
1	.6726	.7258	.8220	.6638	.1494*	.0962*	.0532*	-.0088	-.0620*	-.1582*

Note: * indicates the difference is significant, that is, twice the upper bound of standard error for empirical power ($SE = \sqrt{2 \times \frac{p \times (1-p)}{n}} = .01$ where $p = .5$ and $n = 5000$), $2 \times SE = .02$.

significantly higher than RC. At the small or large effect size, the pairwise power differences among the three VC structures were not all significant across the three tests.

Empirical Power by Sample Size

With the increase of the sample size, the power of TRM, RC AR(1) and UN increased (see Figure 4). It seemed that below the sample size of 100 (i.e., $N = 600$), the four models increased power at an increasing rate and above the sample size of 100, the four enhanced their

power at a decreasing rate, approaching to the asymptote of 1.

When the sample size was small, $n = 25$ (i.e., $N = 150$), the power of all four models was low (e.g., around or below .3 in the treatment and interaction tests). At the sample size of 100 (i.e., $N = 600$), TRM gained power around .75 and .95 in the treatment and time tests, respectively, but below .55 in the interaction test; whereas all three VC structures obtained power merely above .50 and .80 in the treatment and time tests, respectively, but above .60 in the

Table 2. Power Pattern and Pairwise Power Difference of TRM, RC, AR(1) and UN by G Matrix when $n = 75$, $t = 3$ and $\lambda = .5$ of 5000 MC Samples at $\alpha = .10$

G	Power					Power Difference				
	RC	AR(1)	UN	TRM	UN vs. RC	UN vs. AR(1)	AR(1) vs. RC	TRM vs. RC	TRM vs. AR(1)	TRM vs. UN
Two-group Treatment Effect ($\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$)										
$\begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}$	.4382	.4756	.4982	.7020	.0600*	.0226*	.0374*	.2638*	.2264*	.2038*
$\begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$	.4264	.4688	.4946	.6762	.0682*	.0258*	.0424*	.2498*	.2074*	.1816*
$\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$	.4070	.4610	.4918	.6562	.0848*	.0308*	.0540*	.2492*	.1952*	.1644*
Time Effect ($\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$)										
$\begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}$	.7686	.7814	.8002	.9396	.0316*	.0188	.0128	.1710*	.1582*	.1394*
$\begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$	.7354	.7576	.7828	.9284	.0474*	.0252*	.0222*	.1930*	.1708*	.1456*
$\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$	.7144	.7376	.768	.9146	.0536*	.0304*	.0232*	.2002*	.1770*	.1466*
Interaction ($\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = 0.5$)										
$\begin{pmatrix} 1 & .1 \\ .1 & 1 \end{pmatrix}$	.5110	.5318	.5570	.4554	.0460*	.0252*	.0208*	-.0556*	-.0764*	-.1016*
$\begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$	.4844	.5092	.5390	.4368	.0546*	.0298*	.0248*	-.0476*	-.0724*	-.1022*
$\begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix}$	.4618	.4930	.5210	.4132	.0592*	.0280*	.0312*	-.0486*	-.0798*	-.1078*

Figure 3: Power pattern of HLM, AR(1), UN and TRM by effect size in treatment, time and interaction tests when $n = 75$, $t = 3$, $\lambda = .5$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$ of 5000 MC samples at $\alpha = .1$

REPEATED MEASURES AND HMLM IN LONGITUDINAL DATA ANALYSIS

Table 4 Power Pattern and Pairwise Power Difference of RC, AR(1) and UN by Effect Size when $n = 75$, $t = 3$, $\lambda = .5$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$ of 5000 MC Samples at $\alpha = .10$

G	Power					Power Difference				
	RC	AR(1)	UN	TRM	UN vs. RC	UN vs. AR(1)	AR(1) vs. RC	TRM vs. RC	TRM vs. AR(1)	TRM vs. UN
Two-group Treatment Effect ($\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$)										
.20	.1372	.1650	.1746	.2166	.0374*	.0096	.0278*	.0794*	.0516*	.0420*
.50	.4264	.4688	.4946	.6762	.0682*	.0258*	.0424*	.2498*	.2074*	.1816*
.80	.7758	.8042	.8308	.9642	.0550*	.0266*	.0284*	.1884*	.1600*	.1334*
Time Effect ($\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$)										
.20	.2316	.2542	.2698	.3168	.0382*	.0156*	.0226*	.0852*	.0626*	.0470*
.50	.7354	.7576	.7828	.9284	.0474*	.0252*	.0222*	.1930*	.1708*	.1456*
.80	.9804	.9852	.9892	1.0000	.0088	.0040	.0048	.0196*	.0148*	.0108
Interaction ($\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = 0.5$)										
.20	.1686	.1856	.1928	.1602	.0242*	.0072	.0170	-.0084	-.0254*	-.0326*
.50	.4844	.5092	.5390	.4368	.0546*	.0298*	.0248*	-.0476*	-.0724*	-.1022*
.80	.8342	.8498	.8680	.7864	.0338*	.0182	.0156	-.0478*	-.0634*	-.0816*

Note: * indicates the difference is significant, that is, twice the upper bound of standard error for empirical power ($SE = \sqrt{2 \times \frac{p \times (1-p)}{n}} = .01$ where $p = .5$ and $n = 5000$), $2 \times SE = .02$.

Table 5. Power Pattern and Pairwise Power Difference of HLM, AR(1), UN and TRM by Sample Size when $t = 3$, $\beta = .5$, $\lambda = .5$ and $G = \begin{pmatrix} 1 & .3 \\ .3 & 1 \end{pmatrix}$ of 5000 MC Samples at $\alpha = .10$

G	Power				Power Difference					
	RC	AR(1)	UN	TRM	UN vs. RC	UN vs. AR(1)	AR(1) vs. RC	TRM vs. RC	TRM vs. AR(1)	TRM vs. UN
Two-group Treatment Effect ($\beta_{01} = .5, \beta_{10} = 0, \beta_{11} = 0$)										
25	.1900	.2334	.2572	.3368	.0672*	.0238*	.0434*	.1468*	.1034*	.0796*
50	.3166	.3590	.3754	.5270	.0588*	.0164	.0424*	.2104*	.1680*	.1516*
75	.4264	.4688	.4946	.6762	.0682*	.0258*	.0424*	.2498*	.2074*	.1816*
100	.5092	.5502	.5836	.7884	.0744*	.0334*	.0410*	.2792*	.2382*	.2048*
125	.5976	.6362	.6652	.8538	.0676*	.0290*	.0386*	.2562*	.2176*	.1886*
150	.6794	.7130	.7376	.9130	.0582*	.0246*	.0336*	.2336*	.2000*	.1754*
175	.7430	.7750	.7982	.9418	.0552*	.0232*	.0320*	.1988*	.1668*	.1436*
200	.7950	.8228	.8452	.9704	.0502*	.0224*	.0278*	.1754*	.1476*	.1252*
Time Effect ($\beta_{01} = 0, \beta_{10} = .5, \beta_{11} = 0$)										
25	.3790	.4050	.4306	.5320	.0516*	.0256*	.0260*	.1530*	.1270*	.1014*
50	.5970	.6192	.6416	.8058	.0446*	.0224*	.0222*	.2088*	.1866*	.1642*
75	.7354	.7576	.7828	.9284	.0474*	.0252*	.0222*	.1930*	.1708*	.1456*
100	.8504	.8644	.8786	.9740	.0282*	.0142*	.0140	.1236*	.1096*	.0954*
125	.9124	.9230	.9386	.9916	.0262*	.0156*	.0106	.0792*	.0686*	.0530*
150	.9496	.9548	.9658	.9976	.0162*	.0110	.0052	.0480*	.0428*	.0318*
175	.9686	.9736	.9806	.9996	.0120	.0070	.0050	.0310*	.0260*	.0190*
200	.9814	.9842	.986	.9996	.0046	.0018	.0028	.0182*	.0154*	.0136
Interaction ($\beta_{01} = 0, \beta_{10} = 0, \beta_{11} = 0.5$)										
25	.2428	.2638	.2800	.2126	.0372*	.0162	.0210*	-.0302*	-.0512*	-.0674*
50	.3692	.3902	.4166	.3208	.0474*	.0264*	.0210*	-.0484*	-.0694*	-.0958*
75	.4844	.5092	.5390	.4368	.0546*	.0298*	.0248*	-.0476*	-.0724*	-.1022*
100	.5956	.6162	.6422	.5340	.0466*	.0260*	.0206*	-.0616*	-.0822*	-.1082*
125	.6812	.7008	.7264	.6158	.0452*	.0256*	.0196	-.0654*	-.0850*	-.1106*
150	.7442	.7630	.7916	.6896	.0474*	.0286*	.0188	-.0546*	-.0734*	-.1020*
175	.7982	.8146	.8378	.7436	.0396*	.0232*	.0164	-.0546*	-.0710*	-.0942*
200	.8498	.8622	.8838	.8000	.0340*	.0216*	.0124	-.0498*	-.0622*	-.0838*

Note: * indicates the difference is significant, that is, twice the upper bound of standard error for empirical power ($SE = \sqrt{2 \times \frac{p \times (1-p)}{n}} = .01$ where $p = .5$ and $n = 5000$), $2 \times SE = .02$.

REPEATED MEASURES AND HMLM IN LONGITUDINAL DATA ANALYSIS

interaction test. As the sample size reached 200 (i.e., $N = 1200$), the power of all four was above .80 (see Table 5).

Table 5 also showed that the pairwise differences between TRM and the three VC structures were significant across the samples sizes in all three tests. Generally, TRM had significantly higher power than the three VC structures in the treatment and time tests but significantly lower power in the interaction test. It also appeared that the pair-wise differences shrunk as sample sizes increased.

Bootstrap Estimates

The bootstrap estimates, bias, standard errors and standard 95% confidence intervals of the treatment effect were examined within the three factors, G matrix, effect size and sample size per treatment group under specified conditions (see Appendix B). The results indicate that TRM, RC, AR(1) and UN generate unbiased and identical estimates of the treatment effect. TRM has slightly smaller bootstrap standard errors and hence slightly narrower confidence intervals. The bootstrap estimates of all three VC structures have similar patterns within each factor. As the correlation increases, the standard errors become slightly larger and therefore the confidence intervals are wider. As the reliability and sample sizes increase, the bootstrap standard errors decrease and confidence intervals become narrower.

Conclusion

This MC study primarily concerns the empirical power of TRM and HMLM under three variance-covariance (VC) structures in the longitudinal study. Specifically, this paper compared the power of TRM, AR (1) and UN in three tests, two-group treatment effect (β_{01}), time effect (β_{10}) and time-by-treatment interaction (β_{11}), under the balanced design in longitudinal studies. The three factors in this power study are the G matrix (G), reliability (λ), effect size (β) and sample size per treatment group (n).

Researchers have raised the question on what is the power to detect the interactions when they do exist in the HMLM data and expected HMLM perform better than traditional models but without proof (Davison, Kwak, Seo, and Choi, 2002; Kreft, 1996; Raudenbush, 1995). This study provided an empirical power

estimates in the interaction test for both TRM and HMLM. One of the interesting findings in this power study indicates that TRM has significantly lower power than the other three HMLM models, RC, AR(1) and UN, in the interaction test, although it gains the significantly highest power in the main effects tests, treatment and time tests under the balanced design in the specified generic situations.

This study also supplements more comprehensive empirical indexes for estimating the model precision based on the bootstrap estimates and the approximate power for both main effects and interaction tests under more generic situations, including the empirical power indexes of HMLM under three different covariance structures which have not yet been specifically addressed in the literature. Based on this study, TRM could be the choice if researchers are more interested in main effect tests and the practical situation is most similar to this research where the balanced design is assumed and fixed effects are primarily the concern. If researchers are more concerned with interaction tests, this study recommends that UN, AR(1) or RC be the method of choice. When the number of repeated measures is 3, UN has the higher power than AR(1) or RC in the three tests within each factor. UN could be the choice if the practical situation is most similar to this research and if we need to try an exploratory analysis when the VC structure is assumed unknown.

From this study, we noticed that the power can be significantly different among different VC structures when using the HMLM models in the longitudinal study. In addition to referring to the model fit statistics (Akaike, 1973; Littell et al., 2006; Pinheiro & Bates, 2000; Schwarz, 1978; Singer & Willett, 2003), the empirical power results from this study could be a reference source when applying HMLM models. Also from these empirical results, the practitioners may estimate the sample sizes, the reliability, effect size or the correlation in G matrix for their studies if scenarios are similar to this study.

Future studies may consider extending this MC study by comparing power across factors instead of within each factor or fixing conditions and comparing the power by varying

the sample size ratios between the number of subjects and time points while holding the total sample size. Instead of reliability, interclass correlation (ICC) could be considered in the power analysis. Although the magnitude of power difference and power decreasing or increasing rates can vary, the general power patterns among TRM and the three VC structures are expected to be similar to this study. The HMLM data generator and power comparison macro (the author, 2006) could be expanded to generate missing data or non-normal longitudinal data in order to be more practical and to examine the statistical properties and power of more complex growth models.

References

- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-228). Budapest, Hungary: Akademiai Kiado.
- Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101, 1, 147-158.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Mahwah, NJ: Lawrence, Erlbaum Associates.
- Davison, M. L., Kwak, N., Seo, Y. S., & Choi, J. (2002). Using hierarchical linear models to examine moderator effects: Person-by-organization interactions. *Organizational Research Methods*, 5, 231-254.
- 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*, 39(B), 1-8.
- Dempster, A. P., Rubin, D. B., & Tsutakawa, R. K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, 76, 341-353.
- Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, 22, 364-376.
- Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Arnold.
- Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 32, 95-112.
- Hedeker, D., Gibbons, R. D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition: Comparing time-related contrasts between two groups. *Journal of Educational and Behavioral Statistics*, 24(1), 70-93.
- Huynh, H., & Feldt, L.S. (1976). Estimation of the Box Correction for Degrees of Freedom from Sample Data in the Randomized Block and Split Plot Designs. *Journal of Educational Statistics*, 1, 69-82.
- Hox, J. J. (2002). *Multilevel analysis techniques and applications*. Mahwah, N.J.: Lawrence Erlbaum Associates.
- Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, 42, 805-820.
- Kreft, I. G. G. (1996): Are multilevel techniques necessary? An overview, including simulation studies. Retrieved August 13, 2005, <http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html>.
- Littell, R. C., Henry, P. R., & Ammerman, C. B. (1998). Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science*, 76, 1216-1231.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (2006). *SAS system for mixed models*. (2nd ed.). Cary, NC: SAS Institute Inc.
- Little, R. J. A. (1995). Modeling the dropout mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90, 1112-1121.
- Little, R.J., & Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd edition). New York: John Wiley.
- Maas, C. J. M., & Snijders, T. B. (2003). The multilevel approach to repeated measures for complete and incomplete data. *Quality & Quantity*, 37, 71-89.
- McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. New York: John Wiley & Sons, Inc. Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-326.

- Montgomery, D. C. (2005). *Design and analysis of experiments*. Hoboken, NJ : John Wiley & Sons, Inc.
- Muthén, B. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, 29, 81-117.
- Muthén, B. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345-368). Newbury Park, CA: Sage Publications.
- Muthén, L. & Muthén, B. (2006). *Mplus user's guide*. (4th ed.). Los Angeles: Muthén & Muthén.
- Pinheiro, J. C., & Bates, D. M. (2000). *Statistics and computing: Mixed-effects models in S and S-PLUS*. New York: Springer.
- Potthoff, R. F., & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51, 313-326.
- Raudenbush, S. W. (1995). Maximum likelihood estimation for unbalanced multilevel covariance structure models via the EM algorithm. *British Journal of Mathematical and Statistical Psychology*, 48 (2), 359-370.
- Raudenbush S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. London: Sage publications, Inc.
- Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM6: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.
- Raudenbush, S.W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5, 2, 199-213.
- Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2005). *Optimal design for longitudinal and multilevel research: Documentation for the "Optimal Design" software*. Retrieved September 20, 2006, from http://sitemaker.umich.edu/group-based/optimal_design_software
- Rogosa, D. R., & Willett, B. (1985). Understanding correlates of changes by modeling individual differences in growth. *Psychometrika*, 50, 203-228.
- SAS Institute Inc. (2003). *SAS/STAT user's guide, version 9.1*. Cary, NC: SAS Institute Inc..
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis*. New York: Oxford University Press, Inc.
- Skrondal, A. & Rabe-Hesketh, S. (2004). Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. New York: Chapman & Hall/CRC.
- Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B.S. Everitt & D.C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science*. Volume 3, 1570-1573. Chichester: Wiley.
- Schluchter, M. D. (1988). Unbalanced repeated measures models with structured covariance matrices. In W. J. Dixon (Ed.), *BMDP statistical software manual*, 2, 1081-1114. Berkeley, CA: University of California Press.
- Schwarz, G. (1978). Estimating the dimensions of a model. *Annals of Statistics*, 6, 461-464.
- Van der Leeden, R., Vrijburg, K., & de Leeuw, J. (1996). A review of two different approaches for the analysis of growth data using longitudinal mixed linear models: Comparing hierarchical linear regression (ML3,HLM) and repeated measures design with structured covariance matrices (BMDP-5V). *Computational Statistics and Data Analysis*, 21, 5, 583-605.
- Van der Leenden, R. (1998). Multilevel analysis of repeated measures data. *Quality & Quantity*, 32, 15-29.
- Wolfinger, R. D. (1993). Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*, 22, 4, 1079 - 1106.
- Wolfinger, R. D., & Chang, M. (1995). Comparing the SAS GLM and MIXED procedures for repeated measures. *SUGI Proceedings*. Cary, NC: SAS Institute Inc.

Acknowledgement: This research was supported in part by R01 DA014661 from the National Institute on Drug Abuse (Espy, PI), R01 MH065668 from the National Institute of Mental Health (Espy, PI), R01 HD050309 (Taylor, PI) from the National Institute of Child Health and Development. The previous support for the first author by the Voinovich School of Leadership and Public Affairs of Ohio University is also gratefully acknowledged.

Appendix A.

Table 6. Validation of Data Generator Using Potthoff and Roy's Data

	Potthoff and Roy's Data	Simulated Data
Intercept	21.2091	21.2063
Gender	1.4065	1.4065
Time	0.9591	0.9587
Gender*Time	0.6097	0.6115

Appendix B

Table 7. Bootstrap Estimates of Treatment Effect for RC, AR(1), UN and TRM by G Matrix, Effect Size, Reliability and Sample Size of 5000 MC Samples at $\alpha = .10$. Table continued on next page.

(a) RC and RM

		Model									
		RC					RM				
	β_{01}	BIAS	SE	CI_low	CI_high	B01	BIAS	SE	CI_low	CI_high	
G Matrix (ρ)											
0.10	0.50	0.00	0.31	-0.11	1.11	0.50	0.00	0.23	0.06	0.95	
0.30	0.50	0.00	0.31	-0.12	1.12	0.50	0.00	0.23	0.04	0.96	
0.50	0.50	0.00	0.32	-0.12	1.12	0.50	0.00	0.24	0.03	0.97	
Effect size (d)											
0.20	0.20	0.00	0.31	-0.42	0.82	0.20	0.00	0.23	-0.26	0.66	
0.50	0.50	0.00	0.31	-0.12	1.12	0.50	0.00	0.23	0.04	0.96	
0.80	0.80	0.00	0.31	0.18	1.42	0.80	0.00	0.23	0.34	1.26	
Reliability (λ)											
0.01	0.49	-0.01	2.59	-4.59	5.58	0.49	-0.01	1.62	-2.68	3.66	
0.25	0.50	0.00	0.48	-0.45	1.45	0.50	0.00	0.33	-0.14	1.14	
0.50	0.50	0.00	0.31	-0.12	1.12	0.50	0.00	0.23	0.04	0.96	
0.75	0.50	0.00	0.23	0.05	0.95	0.50	0.00	0.19	0.12	0.88	
1.00	0.50	0.00	0.17	0.16	0.84	0.50	0.00	0.17	0.16	0.84	
Sample Size (n)											
25	0.50	0.00	0.54	-0.56	1.56	0.50	0.00	0.41	-0.31	1.31	
50	0.50	0.00	0.38	-0.24	1.24	0.50	0.00	0.29	-0.06	1.06	
75	0.50	0.00	0.31	-0.12	1.12	0.50	0.00	0.23	0.04	0.96	
100	0.50	0.00	0.27	-0.04	1.03	0.50	0.00	0.20	0.10	0.90	
125	0.50	0.00	0.25	0.02	0.98	0.50	0.00	0.18	0.14	0.86	
150	0.50	0.00	0.23	0.06	0.94	0.50	0.00	0.17	0.17	0.83	
175	0.50	0.00	0.20	0.10	0.90	0.50	0.00	0.16	0.19	0.81	
200	0.50	0.00	0.19	0.12	0.88	0.50	0.00	0.15	0.22	0.79	

(b) AR(1) and UN										
	Model									
	AR(1)					UN				
	β_{01}	BIAS	SE	CI_low	CI_high	B01	BIAS	SE	CI_low	CI_high
G Matrix (ρ)										
0.10	0.50	0.00	0.31	-0.12	1.12	0.50	0.00	0.31	-0.11	1.11
0.30	0.50	0.00	0.32	-0.12	1.12	0.50	0.00	0.31	-0.11	1.11
0.50	0.50	0.00	0.32	-0.13	1.13	0.50	0.00	0.31	-0.11	1.11
Effect Sizes (d)										
0.20	0.20	0.00	0.32	-0.42	0.82	0.20	0.00	0.31	-0.41	0.81
0.50	0.50	0.00	0.32	-0.12	1.12	0.50	0.00	0.31	-0.11	1.11
0.80	0.80	0.00	0.32	0.18	1.42	0.80	0.00	0.31	0.19	1.41
Reliability (λ)										
0.01	0.49	-0.01	2.60	-4.60	5.58	0.49	-0.01	2.60	-4.60	5.59
0.25	0.50	0.00	0.49	-0.45	1.45	0.50	0.00	0.48	-0.45	1.45
0.50	0.50	0.00	0.32	-0.12	1.12	0.50	0.00	0.31	-0.11	1.11
0.75	0.50	0.00	0.24	0.04	0.96	0.50	0.00	0.22	0.06	0.93
1.00	0.50	0.00	0.18	0.14	0.86	0.50	0.00	0.16	0.19	0.81
Sample Size (n)										
25	0.50	0.00	0.55	-0.58	1.57	0.50	0.00	0.54	-0.56	1.56
50	0.50	0.00	0.38	-0.25	1.25	0.50	0.00	0.37	-0.24	1.23
75	0.50	0.00	0.32	-0.12	1.12	0.50	0.00	0.31	-0.11	1.11
100	0.50	0.00	0.28	-0.04	1.04	0.50	0.00	0.27	-0.03	1.02
125	0.50	0.00	0.25	0.01	0.98	0.50	0.00	0.24	0.02	0.97
150	0.50	0.00	0.23	0.05	0.95	0.50	0.00	0.22	0.06	0.93
175	0.50	0.00	0.21	0.10	0.90	0.50	0.00	0.20	0.11	0.89
200	0.50	0.00	0.19	0.12	0.88	0.50	0.00	0.19	0.13	0.87

Comparing Different Methods for Multiple Testing in Reaction Time Data

Massimiliano Pastore Massimo Nucci Giovanni Galfano
Università di Padova

Reaction times were simulated for examining the power of six methods for multiple testing, as a function of sample size and departures from normality. Power estimates were low for all methods for non-normal distributions. With normal distributions, even for small sample sizes, satisfactory power estimates were observed, especially for FDR-based procedures.

Keywords: multiple testing, reaction times, power, False Discovery Rate, Type I error

Introduction

Scientific research often deals with the problem of performing many tests of significance. However, this practice results in an increase of the likelihood of committing one or more Type I errors, which grows as the number of tests increases (e.g., Keselman, Cribbie, & Holland, 1999). In the most common approach, error rate is familywise controlled (Familywise Error Rate, FWER) by reducing the α value as a direct function of the number of comparisons to be computed. In the classic Bonferroni method (1936), the threshold probability (α_{FW} , usually set at .05) is divided by the total number of comparisons. This approach to controlling errors in multiple-testing contexts ensures that the probability of committing Type I error at least

once is $\leq \alpha$. The intrinsic limit of multiple testing with FWER control is that such approach becomes more conservative as the number of tests rises: Indeed, a major criticism frequently levelled at multiple testing is their lack of power.

A different perspective to controlling Type I error when performing many tests of significance is represented by the False Discovery Rate (FDR). This statistical procedure, introduced by Benjamini and Hochberg (1995), can be implemented in all those experimental contexts in which the computation of a large number of comparisons is required. The FDR is focused on the proportion of errors committed when H_0 is rejected, which results in both keeping Type I error under control and in an increase of power. Further advantages characterizing FDR are represented by its easy and quick implementation (Thissen, Steinberg, & Kuang, 2002), and by its wide applicability, as proved by the fact that FDR can be adopted when multiple comparisons involve either independent or correlated test statistics (Benjamini & Yekutieli, 2001).

A third possible way for dealing with multiple testing is represented by resampling-based procedures (Westfall & Young, 1993). Following this approach, the values of observed variables are randomly re-assigned to the experimental groups, and then the test statistics are re-computed. Thus, the resampling-based p-value is the proportion of resampled data sets yielding a statistic as extreme as the original statistic.

Massimiliano Pastore is Assistant Professor of Psychometrics in the Faculty of Statistical Science, Dipartimento di Psicologia dello Sviluppo e della Socializzazione, at the University of Padova. E-mail: massimiliano.pastore@unipd.it. Massimo Nucci is Assistant Professor of Psychometrics in the Faculty of Psychology, Dipartimento di Psicologia Generale at the University of Padova. Giovanni Galfano is Assistant Professor of Experimental Psychology in the Faculty of Psychology, Dipartimento di Psicologia dello Sviluppo e della Socializzazione at the University of Padova.

In this article, a Monte Carlo study is illustrated aimed at comparing the performance of six different procedures for treating multiple testing. The analysis has been conducted in the context of multiple comparisons among means resulting from nonnormally-distributed and correlated variables. Specifically, the classic Bonferroni method, two single-step FDR methods, two resampling-based methods, and a combined resampling-based FDR procedure were examined.

These methods were used for adjusting p-value and then comparing their power. Because in multiple comparison testing more than one definition of power and Type I error rate is available, three different definitions associated to these measures were considered. It should be noted that FDR-based methods do not control for FWER (e.g., Wilcox, 2003). However, it is important to remark that comparing power of methods that do not have similar control over Type I errors can provide critical information as to the choice of a particular test in light of the associated costs (in terms of Type I error) and benefits (in terms of power; e.g., Horn & Dunnett, 2004).

Reaction time (RT) data were simulated for this research. The present study focused on this particular type of variable for two main reasons. First, RTs represent the dominant dependent measure in cognitive psychology (e.g., Van Zandt, 2002). Second, RTs possess critical features that make them hard to be analyzed with classical statistical procedures (Heathcote, 1996).

In the most common experimental paradigms using RTs, participants are submitted to a series of stimuli that have to be responded to as fast as possible. Therefore, measurements can hardly be considered as independent from each other. In addition, it is well known that RTs are not distributed according to a normal function (e.g., Schwarz, 2001; Van Zandt, 2000). McGill (1963) and Hohle (1965) proposed as a descriptive model of RTs, a theoretical distribution obtained through the convolution of a normal distribution and an exponential distribution, subsequently known as ex-Gaussian (Burbeck & Luce, 1982). Although other descriptive models are available such as the ex-Wald, the Weibull and the Gamma distributions

(see, e.g., Schwarz, 2001; Van Zandt, 2000), to date the ex-Gaussian distribution is among the most representative models for describing RTs (Ratcliff, 1978; 1979; Ratcliff & Murdock, 1976). In addition, it is worth noting that, using the ex-Gaussian model, the usefulness of decomposing the normal and exponential components has been consistently demonstrated (e.g., Heathcote, 1996; Heathcote, Popiel, & Mewhort, 1991). For example, the simple arithmetic mean cannot be considered a satisfactory statistic within this context, given the skewness characterizing RTs. By contrast, there is wide agreement that ex-Gaussian parameters are more appropriate for describing (and interpreting) RTs (Heathcote, 1996). In the present paper, the ex-Gaussian distribution was adopted as a plausible model for RT data.

An experimental setting with three stimuli requiring a response of some sort was simulated. Each stimulus was repeated three times. Multiple comparisons among the observed RT means, obtained in this hypothetical task were then performed. Both sample size and the magnitude of the RT exponential component, were manipulated. The estimated power of the six procedures was then compared. Before illustrating the methods and results of the Monte Carlo study, the basics of p-value adjustment in the examined procedures will be outlined, and the features of the ex-Gaussian distribution and analysis will be briefly summarized.

p-Value Adjustment

Suppose there is interest in testing m hypotheses simultaneously. For each hypothesis H_i , $i=1,2,\dots,m$, m test statistics and the relative p-values will be computed. It is possible to compute an adjusted p-value $\tilde{p}_i$ for each test. Thus, the decision to reject H_i at $\text{FWER} = \alpha$ is obtained by merely checking whether $\tilde{p}_i \leq \alpha$. According to Westfall and Young (1993, p. 11), the mathematical definition of an adjusted p-value is as follows:

$$\tilde{p}_i = \inf\{\alpha : H_i \text{ is rejected at FWER} = \alpha\} \quad (1)$$

That is, $\tilde{p}_i$ is the smallest significance level for which one still rejects H_i , given a particular simultaneous test procedure. Adjusted p-values for FDR controlling procedures are defined similarly (Yekutieli & Benjamini, 1999):

$$\tilde{p}_i = \inf\{\alpha : H_i \text{ is rejected at FDR} = \alpha\} \quad (2)$$

In the present study, the following p-value adjustment procedures were considered: Bonferroni adjustment (B), two single-step FDR-type adjustments, that is Benjamini-Hochberg (BH; the basic FDR method) and Benjamini-Yekutieli (BY), and three resampling-based adjustments, that is the method described by Reiner, Yekutieli and Benjamini (2003; RYB) and two methods proposed by Ge, Dudoit and Speed (2003), called maxT and minP. Whereas B, minP, and maxT control FWER, BH, BY, and RYB control FDR.

B adjustment

This adjustment by Bonferroni (1936) consists of multiplying each observed probability, p_i , by the number of comparisons that have been performed. In case the value resulting from this computation exceeds 1, then probability is set at 1:

$$_B \tilde{p}_i = \min(p_i m, 1) \quad (3)$$

BH adjustment

This method has been introduced by Benjamini and Hochberg (1995) for independent and positive regression dependent test statistics. Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the observed probabilities arranged in increasing order, then:

$$_{BH} \tilde{p}_i = \min \left\{ p_{(j)} \frac{m}{j} : i \leq j \right\}; \quad j = 1, \dots, m \quad (4)$$

BY adjustment

This method was proposed by Benjamini and Yekutieli (2001) for controlling general dependency structures. Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the observed probabilities arranged in increasing order, then:

$$_{BY} \tilde{p}_i = \min \left\{ p_{(j)} \frac{m \sum_{k=1}^m \frac{1}{k}}{j} : i \leq j \right\}; \quad j = 1, \dots, m \quad (5)$$

RYB adjustment

This is a resampling-based FDR adjustment. In particular the method described by Reiner et al. (2003) was considered, which can be summarized as follows: First, the data are repeatedly resampled under complete null hypothesis (meaning that all H_i are true) and a vector of resampling-based p-values is computed for each H_i . For the k-th hypothesis, with an observed test statistics t_k , the estimated p-value is:

$$p_k^{est} = \frac{\#\{t_{ij}^* : |t_{ij}^*| \geq |t_k|\}}{mN} \quad (6)$$

where $i = 1, \dots, m$, m are the number of hypotheses, $j = 1, \dots, N$, N the number of resampling, and t_{ij}^* are the resampling-based test statistics.

The adjusted p-values using the BH adjustment is obtained as follows:

$$_{RYB} \tilde{p}_i = \min \left\{ p_{(k)}^{est} \frac{m}{k} : i \leq k \right\} \quad (7)$$

Resampling maxT adjustment

This algorithm, originally proposed by Westfall and Young (1993), has been further examined by Ge et al. (2003). The step-down maxT adjusted p-values are defined by:

$$\max_T \tilde{p}_i = \max_{k=1, \dots, i} \{ \Pr(\max_{l=k, \dots, m} |T_l| \geq |t_{(k)}| / \text{under complete null hypothesis}) \} \quad (8)$$

where $|T_l|$ is the random variable associated to the statistical test, and $|t_{(1)}| \geq |t_{(2)}| \geq \dots \geq |t_{(m)}|$ denote the ordered observed test statistics.

Resampling minP adjustment

This algorithm was also put forward by Westfall and Young (1993). However, the version considered in the present study is based on a modified adaptation (see Ge et al., 2003). The step-down minP adjusted p-values are defined by:

$$\min_P \tilde{p}_i = \max_{k=1, \dots, i} \{ \Pr(\min_{l=k, \dots, m} P_l \leq p_{(k)} / \text{under complete null hypothesis}) \} \quad (9)$$

where P_l denotes the random variable for the unadjusted p-value of the l -th hypothesis and $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ denote the ordered observed p-values.

Ex-Gaussian Distribution

The ex-Gaussian function is identified as a good theoretical approximation of RT distribution (e.g., Heathcote, 1996; Heathcote et al., 1991; Van Zandt, 2000) and its shape can be formally described as follows:

$$f(t) = \frac{1}{t} e^{\frac{t}{\tau} + \frac{\mu}{\tau} + \frac{\sigma^2}{2\tau^2}} \Phi \left(\frac{\tau - \mu - \frac{\sigma^2}{\tau}}{\sigma} \right) \quad (10)$$

where

$$\Phi(z) = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz \quad (11)$$

This density function depends on three parameters: μ and σ , corresponding to mean and standard deviation of the Gaussian component of the distribution respectively, and τ , corresponding to the exponential component

of the distribution. Ratcliff (1979) showed that mean (${}_{RT}\mu$) and standard deviation (${}_{RT}\sigma$) of the ex-Gaussian can be rewritten as a function of these three parameters. In particular:

$${}_{RT}\mu = \mu + \tau \quad (12)$$

and

$${}_{RT}\sigma = \sqrt{\sigma^2 + \tau^2} \quad (13)$$

Examples of ex-Gaussian density functions are depicted in Figure 1, where the influence of the exponential component on the shape of the distribution function is illustrated. The curves have $\mu = 550$ and $\sigma = 50$ as fixed parameters, whereas the τ value is varied. It is worth noticing that the exponential component determines an increase of the positive skew.

As briefly anticipated earlier, Heathcote (1996; Heathcote et al., 1991; Mewhort, Braun, & Heathcote, 1992) has proposed an RT analysis method based on the properties highlighted above. In particular, Heathcote (1996) has developed a statistical package, RTSYS, that allows researchers to easily obtain values for μ , σ , and τ by means of RT decomposition.

Several arguments support the need of using an RT decomposition technique prior to statistical analysis. First, RT data can contain extreme values (i.e., outliers) that do not reflect the effects of the independent variables and can be problematic for interpreting the results. Solutions to the problem of outliers usually rely on trimming observations (e.g., Ratcliff, 1993; Van Selst & Jolicoeur, 1994; Wilcox, 2005). However, finding a general criterion for removing data is problematic because real data are almost inevitably rejected along with spurious data. Second, as discussed above, skew in RT distribution can cause serious problems of interpretation for descriptive statistics. For instance, a given independent variable may influence the mean and median differently by modifying the degree of skew. It should also be stressed that significantly skewed data violate the assumption underlying most parametric tests, that variability in data is normal. Whereas the common approach in research practice is to

ignore skew, several studies have shown that the magnitude of skew often contains information about the effect of experimental manipulations (Ratcliff & Murdock, 1976; Heathcote et al., 1991; Campbell & Penner-Wilger, 2006). It follows that even if one circumvents the problem of violating the normality assumption of parametric tests by transforming RTs, the risk of losing information and missing potentially important effects is still present.

In summary, through quantifying RT distribution shape, ex-Gaussian decomposition can reveal structure within RT data not revealed by conventional analyses. It has successfully been adopted in a variety of studies dealing with RTs in several research fields related to cognitive psychology (e.g., Andrews & Heathcote, 2001; Armstrong & Munoz, 2003; Balota & Spieler, 1999; Dell'Acqua, Job, & Grainger, 2001; Leth-Steensen, Elbaz, & Douglas, 2000; Madden, Gottlob, Denny, Turkington, Provenzale, Hawk, & Coleman, 1999; Mewhort, & Johns, 2000; Penner-Wilger, Leth-Steensen, & LeFevre, 2002; Spieler, Balota, & Faust, 2000; West, Murphy, Armilio, Craik, & Stuss, 2002).

Methodology

Data Generation

Data were generated and analyzed by means of a custom-made program written in R (R Development Core Team, 2003). Random number generation was achieved by using the Mersenne-Twister method (Matsumoto & Nishimura, 1998). This generator guarantees far longer period and far higher order of equidistribution than any other implemented generators.

RTs were generated through the application of the `rnorm` function concerning the normal component (with μ and σ as mean and standard deviation, respectively) and the `rexp` function concerning the exponential component (with τ as parameter), as follows:

$$RT = rnorm(n, \mu, \sigma) + rexp(n, 1/\tau) \quad (14)$$

Clearly, with $\tau = 0$, the exponential component is set to 0. As a result, the ex-Gaussian function reduces to a normal distribution with mean μ and standard

Figure 1. Ex-Gaussian density functions with $\mu = 550$, $\sigma = 50$ and $\tau \in \{0, 100, 200, 300\}$

deviation σ . Note that this very same values would be obtained after performing the RT decomposition algorithm (Heathcote, 1996) in any generated RT raw data set.

In order to generate correlated data, the method described by Jöreskog and Sörbom (1996, pp. 189-190) was used. Such method is based on the adoption of a lower triangular matrix T such that $\Sigma = TT'$, where Σ is the population correlation matrix. Application of such procedure ensures the generation of correlated ex-Gaussian distributions.

Experimental Design

An experimental situation with three stimuli (e.g., pictures) requiring a speeded response of some sort in a given task (e.g., picture naming) was simulated. Each stimulus was repeated three times. Multiple comparisons were then performed among all the observed RT means. In such a context, differences may be expected for comparisons between different stimuli. Conversely, no differences should be expected in comparisons between repetitions of the same stimulus. Whereas in an empirical setting this latter type of comparisons may be relevant to test the consistency of a given stimulus (or participant), in the present study it was critical for evaluating Type I error.

The parameters for the simulation were chosen after an extensive review and analysis of the studies employing the ex-Gaussian decomposition technique cited above. Specifically, RT means ranged from 446 (Spieler et al., 2000) to 1199 milliseconds (Leth-Steensen et al., 2000). Using the ex-Gaussian decomposition, the mean value of μ was about 522 milliseconds, ranging from 286 (Dell'Acqua et al., 2001) to 865 (Leth-Steensen et al., 2000). σ varies between 32 (Spieler et al., 2000) and 175 (Leth-Steensen et al., 2000), with mode 50. The estimated values of τ ranged from 41 (Spieler et al., 2000) to 414 (Leth-Steensen et al., 2000). Consequently, three distributions were considered (one for each of the three stimuli) with mean $\mu_1 = 595$, $\mu_2 = 550$, and $\mu_3 = 535$, all of which had a standard deviation of $\sigma = 50$, and four values of τ : 0, 100, 200, and 300. In addition, the correlation value across distributions was set to $\rho = .6$, with the purpose

of simulating a setting with a medium-to-high correlation level, and the correlation value within distributions was set to $\rho = 1$.

To summarize, the notation $ExG(\mu, \sigma, \tau)$ was used to indicate a generic ex-Gaussian distribution with μ , σ , and τ as parameters. Consequently, the resulting three distributions were defined as follows:

$$D_1 \approx ExG(595, 50, \tau)$$

$$D_2 \approx ExG(550, 50, \tau)$$

$$D_3 \approx ExG(535, 50, \tau)$$

$$\text{where } \tau \in \{0, 100, 200, 300\}.$$

The manipulation of τ was aimed to evaluate the performance of the six p -value adjustment methods as a function of departures from normality.

For each of the three distributions (D_1 , D_2 and D_3) three repetitions were performed, thus producing nine RTs in total. A scheme representing the procedure adopted is depicted in Figure 2. The sample size was varied in four different sizes (n): 12, 20, 40, and 80. These particular values were chosen because they are representative of those generally adopted in empirical research (e.g., Andrews & Heathcote, 2001; Dell'Acqua et al., 2001).

By combining the four chosen τ values with the four different sample sizes, sixteen different scenarios were obtained. For each scenario, the sampling was replicated five thousand times. Therefore, the total number of generated samples was $4 \times 4 \times 5000 = 80000$.

Pairwise Comparisons

For each sample, after computing mean RTs, all the possible paired comparisons were performed by means of paired samples t -tests,

$$\text{equals to } \binom{9}{2} = 36. \text{ In order to determine}$$

whether the difference was statistically significant, the p -value adjustments described earlier were used:

MULTIPLE TESTING IN REACTION TIME DATA

Figure 2. Schematic summary of data generation.
 X_{ij} with $i, j \in \{1, 2, 3\}$, is the j -th variable obtained from the D_i distribution.

1. B: following the procedure specified in (3).
2. BH: following the procedure specified in (4).
3. BY: following the procedure specified in (5).
4. RYB: following Resampling FDR Adjustment definition described in (6) and (7); a modified version of the *R* program by A. Reiner available over the internet at <http://www.math.tau.ac.il/~ybenja> was used. For each of the 5000-generated raw-data sets, data were resampled 1000 times.
5. maxT: following definition (8). For each of the 5000-generated raw data sets, data were resampled 1000 times.
6. minP: following definition (9). For each of the 5000-generated raw data sets, data were resampled 1000 times.

For both maxT and minP, the R-package Multtest by Dudoit and Ge was used. This may be downloaded from the Bioconductor website <http://www.bioconductor.org/>.

The tested hypothesis was the following:

$$H_{0:RT} \mu_{ij} - \mu_{hk} = 0$$

where

$$i, j, h, k \in \{1, 2, 3\}$$

and

$$(i, j) \neq (h, k) \quad (15)$$

This hypothesis is true when the comparison is made between two variables belonging to the same distribution, and false when the variables belong to different distributions. The Null Hypothesis status for the considered comparisons is shown in Table 1.

The true values of the differences between means ($\theta = \mu_{ij} - \mu_{hk}$), are represented in Table 2. As a result, nine comparisons for each of the θ values were considered. Note that when $\theta = 0$, H_0 is true, being false in all the other cases.

Table 1: Null Hypothesis status in the examined comparisons

	μ_{11}	μ_{12}	μ_{13}	μ_{21}	μ_{22}	μ_{23}	μ_{31}	μ_{32}
μ_{12}	true							
μ_{13}	true	true						
μ_{21}	false	false	false					
μ_{22}	false	false	false	true				
μ_{23}	false	false	false	true	true			
μ_{31}	false	false	false	false	false	false		
μ_{32}	false	false	false	false	false	false	true	
μ_{33}	false	false	false	false	false	false	true	true

Table 2. True value of the differences between means

Comparisons	θ	
$\mu_{ij} - \mu_{ik}$	0	with $i, j, k \in \{1,2,3\}$ for $j \neq k$
$\mu_{2j} - \mu_{3k}$	15	with $i, j, k \in \{1,2,3\}$
$\mu_{1j} - \mu_{2k}$	45	with $i, j, k \in \{1,2,3\}$
$\mu_{1j} - \mu_{3k}$	60	with $i, j, k \in \{1,2,3\}$

Empirical Evaluation of Power and Type I Error Rate

Because the present study was aimed at evaluating the power of each adjustment procedure, defining power represents a critical issue. Crucially, in multiple testing situations, power is not univocally characterized. the present study, three types of power were considered: Any-pair power was the probability of correctly rejecting at least one hypothesis for each level of $\theta > 0$ (Ramsey, 1978; Westfall & Young, 1993, p. 205). Consequently, the number of times, for each level of $\theta > 0$, in which H_0 was rejected

at least once was computed. This value was then divided by the total number of replications (i.e., 5000). In the experimental practice, the any-pair definition is generally chosen for dealing with exploratory scenarios, because of a higher discriminatory capability.

All-pair power was the probability of correctly rejecting all hypotheses for each level of $\theta > 0$ (Ramsey, 1978; Westfall & Young, 1993, p. 205). Consequently, the number of times, for each level of θ , in which all H_0 were rejected was computed. This value was then divided by the total number of replications

(i.e., 5000). In the experimental practice, the all-pair power definition is generally chosen when missing the rejection of even a single false H_0 has particularly dramatic consequences.

Per-pair power was the rejection probability for a given pair of hypotheses, for each level of $\theta > 0$ (Ramsey, 1978). Consequently, the number of rejected H_0 was counted and then divided by the total number of hypotheses for each level of $\theta > 0$ (i.e., $9 \times 5000 = 45000$). In the experimental practice, the per-pair power definition is generally adopted in meta-analytic contexts (Westfall & Young, 1993), and can be interpreted as an intermediate solution between any-pair and all-pair definitions.

When $\theta = 0$, all H_0 are true. Hence, the number of times in which H_0 were rejected was evaluated for estimating Type I error rate. Three types of Type I error rate were considered:

FWER was the probability of rejecting at least one true null hypothesis. Consequently, the number of times in which H_0 was rejected at least once was counted. This value was then divided by the total number of replications (i.e., 5000).

FDR was the expectation of the proportion of the rejected null hypotheses which are erroneously rejected. Consequently, the proportion of erroneously rejected H_0 was counted. This value was then divided by the total number of replications (i.e., 5000).

Per-Comparison error rate (PCER) was the rejection probability for a given pair of true null hypotheses. Consequently, the number of rejected H_0 was counted and then divided by the total number of hypotheses in which $\theta = 0$ (i.e., $9 \times 5000 = 45000$).

Because the computed values associated to the different power and Type I error definitions vary as a function of the proportion of true null hypotheses (cfr. Dudoit, Shaffer, & Boldrick, 2003), it is worth noting that, in the present context, this proportion was .25.

Results

For each of the sixteen considered scenarios, before estimating power, the mean number of significant tests for all the considered values of the θ parameter was computed. It must be stressed once again that sampling was replicated five thousand times.

Type I error rates

Type I error estimates are illustrated in Table 3. Given that the different methods control different kinds of Type I error, following Dudoit et al. (2003), FWER estimates are reported for B, maxT and minP, whereas FDR estimates are reported for BH, BY and RYB. In addition, PCER estimates are reported for unadjusted p-values (rawp). Inspection of Table 3 shows that B always succeeded in keeping Type I error under .05. The performance of all the remaining methods was modulated by both sample size (n) and the magnitude of the exponential component (τ). More specifically, all methods were weakened as τ increased, whereas increasing sample size resulted in a more efficient control. Crucially, however, when sample size was sufficiently large ($n = 80$), all the FDR-based methods (BH, BY, and RYB) were effective in controlling Type I error adequately even when the magnitude of the exponential component was highest ($\tau = 300$).

Any-Pair Power

Figures 3 and 4 represent the power estimates obtained with n set at 12 and 80, respectively. The four graphs in each figure represent the functions obtained for each specific τ value (0, 100, 200, 300) with the six different methods. In abscissa the value of the θ parameter (i.e., the real difference between means) is represented.

As a general trend, an expected increase of significant results as both θ and n increased can be observed. However, it is worth remarking that the number of significant tests dramatically decreased as τ increased, thus showing that departures from normality directly result in a loss of power.

For $\tau > 0$, RYB showed the best performance when sample size was small ($n = 12$). As sample size increased, however, RYB

performance was equivalent to BH performance in all conditions. When sample size was sufficiently large ($n = 80$), all methods seemed to achieve a good performance in terms of power even for moderate values of the exponential component ($\tau = 100$). Finally, for $\tau > 100$, RYB and BH showed the best performance, followed by maxT and minP (showing overlapping functions), and BY and B (showing overlapping functions).

Per-Pair Power. Figures 7 and 8 illustrate the power estimates for Per-pair definition. This way of defining power results in estimated values that occupy an intermediate level in between Any-pair and All-pair definitions. The results showed similar patterns, whereby power was influenced by both sample size and the magnitude of the exponential component. In more detail, for $n = 12$ none of the methods achieved reasonable power levels. Moreover, for $n = 80$, the methods showed acceptable power levels only for . In good agreement with the results emerged for the previous power definitions, RYB and BH resulted the best adjustment methods, followed by maxT, minP and BY, and B.

In general, the results seem to suggest that for small sample sizes (e.g., $n = 12$, Figures 3, 5 and 7) the power of all methods tended to lower as the value of τ increases, meaning that the likelihood of committing a Type II error tends to rise as the distribution progressively departs from normality. The performance of RYB and BH always proved the best. Also, a general order relationship emerged, for every power definition, so that

$$RYB \cong BH \geq \max T \cong \min P \geq BY \cong B \quad (16)$$

where $X \cong Y$ denotes that X is approximately equivalent to Y , and $\geq$ denotes that X is equivalent or more powerful than Y .

For $n = 80$ (Figures 4, 6 and 8), all methods achieved acceptable power estimates even when $\tau = 100$, provided that $\theta > 15$. This seems to suggest that, with a large sample size, departures from normality do not strongly affect power. When $\tau = 200$, neither RYB nor BH revealed a fully satisfactory performance even for $\theta > 45$. These two methods tended to

produce an equivalent performance in all the different scenarios. The order relationship emerged in the situations with lower sample sizes was confirmed, with RYB and BH being the most powerful methods, and B the least.

Conclusion

The present article was aimed at comparing the power of six different p-value adjustment procedures for treating multiple testing. In particular, RTs, which are the main dependent variable in many experimental contexts related to cognitive psychology (Van Zandt, 2002), were considered. Because it is well known that RTs are not distributed normally, the six p-value adjustment procedures were evaluated by manipulating the parameters related to the Ex-Gaussian distribution. This distribution was chosen because it is one of the most prominent descriptive models for RTs in the literature (Van Zandt, 2000). In order to maintain a close reference with empirical research, the values of the different parameters were chosen based on a series of studies that have employed an RT decomposition technique. This allowed for the examination of the effects of departures from normality on the power estimate associated to each different p-value adjustment procedure. In addition, sample size was manipulated, whose values were selected following the same studies that used the RT decomposition technique. Because sample size is often quite small, the present study tested whether this factor played a major role in modulating the shape of the power function.

As a general comment, two main results emerged in the present investigation. First, the power of the different adjustment procedures was substantially influenced by both sample size and the shape of the distribution. Second, the adjustment procedures included in the present study can be ordered in a constant relationship. In particular, RYB always resulted the most powerful method, although closely followed by BH, whereas B, as expected, appeared very conservative in all the different scenarios. The difference between the most powerful methods (i.e., RYB and BH) and the remaining adjustment procedures was more pronounced for $\theta = 15$. This result is important, because the

MULTIPLE TESTING IN REACTION TIME DATA

Table 3. Type I error estimates as a function of sample size (n) and departures from normality (τ). FWER estimates are reported for B, maxT and minP, FDR estimates are reported for BH, BY and RYB, and PCER estimates are reported for unadjusted p-values (rawp).

n	τ	PCER		FWER		FDR		
		rawp	B	maxT	minP	BH	BY	RYB
12	0	.048	.011	.035	.033	.012	.004	.012
	100	.062	.026	.068	.051	.054	.025	.060
	200	.069	.036	.100	.073	.142	.105	.200
	300	.070	.045	.112	.081	.255	.193	.327
20	0	.049	.011	.038	.038	.012	.003	.012
	100	.056	.029	.067	.061	.031	.017	.032
	200	.065	.045	.099	.083	.112	.096	.138
	300	.067	.047	.099	.083	.216	.206	.248
40	0	.048	.010	.043	.044	.012	.003	.012
	100	.055	.026	.060	.056	.016	.006	.016
	200	.057	.033	.073	.065	.046	.037	.051
	300	.060	.037	.072	.063	.106	.094	.108
80	0	.050	.010	.042	.042	.013	.003	.013
	100	.052	.020	.054	.052	.013	.004	.013
	200	.053	.027	.060	.058	.023	.012	.023
	300	.054	.025	.056	.054	.048	.029	.050

Figure 3. Any-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). B = Bonferroni method (FWER); BH = Benjamini-Hochberg (FDR); BY = Benjamini-Yekutieli (FDR); RYB = Reiner-Yekutieli-Benjamini (resampling-based FDR); minP and maxT (resampling). The horizontal line refers to .05.

MULTIPLE TESTING IN REACTION TIME DATA

Figure 4. Any-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

Figure 5. All-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

MULTIPLE TESTING IN REACTION TIME DATA

Figure 6 All-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

Figure 7. Per-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 12$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

FIGURE 8 Per-pair power estimates for the different p-value adjustment methods as a function of the true difference between means (θ) for $n = 80$. Each graph refers to a different τ value (0 to 300, from left to right). Conventions as in Figure 3.

phenomena investigated in cognitive psychology and mental chronometry are often inferred on the grounds of mean differences in similar orders of magnitude. Relevant examples are offered by the Simon effect (see Lu & Proctor, 1995, for a review), the inhibition of return effect (see Klein, 2000, for a review), and the semantic priming effect (see Neely, 1991, for a review).

In more detail, several observations can be made related to the different controlling methods. Within the class of FWER controlling methods, as illustrated in Table 3, minP and maxT showed a good Type I error control only when $\tau = 0$. For $\tau > 0$, Type I error was not controlled anymore, although it can be observed that performance in this regard increased as n increased. On the other side, minP and maxT showed a clearly higher performance in terms of

power, for small sample sizes, provided that $\tau \leq 100$ (see Figures 3, 5, and 7). With large sample sizes and $\tau = 0$, particularly when $\theta \geq 45$, minP, maxT and B showed overlapping power functions (see Figures 4, 6, and 8). In light of these arguments, minP and maxT may be preferred in the former scenario, whereas B is certainly to be preferred in the latter scenario. Notably, these results hold for all the different power types. Within the class of FDR controlling methods, Table 3 inspection highlights that all methods showed a good Type I error control when $\tau = 0$. Surprisingly, some sort of linear relation seems to characterise Type I error control as a function of n and τ . In particular, when $n = 20$, all methods controlled Type I error for $\tau \leq 100$. When $n = 40$, Type I error control was extended to $\tau = 200$, and when

$n = 80$, a good Type I error control was observed even for $\tau = 300$. In terms of power, BH and RYB consistently showed a better performance than BY, across all conditions and power types (see Figures 3-8). Only for $n = 12$ and $\tau \geq 100$, RYB behaved slightly better than BH, independently of power type. In all the other conditions, the BH method is recommended, because of its quick and easy implementation (Thissen et al., 2002).

When comparing methods controlling a different kind of Type I error, several observations can be made. First, with $\tau = 0$, B should be preferred over FDR-based methods when $\theta \geq 45$ and $n = 80$. In fact, given that they show overlapping power estimates, it may seem more reasonable to choose the method providing the strongest Type I error control. By contrast, when BH and RYB show a clear power advantage over B (e.g., for $n = 12$ and $\tau = 0$), it may be more appropriate choosing either of these FDR-based control methods.

In general, the RT exponential component produced a conspicuous loss of power, especially when sample size was small. For $\tau = 300$, no method among those included in the present study showed power estimates higher than .4, even when the real distance among means was 60 and $n = 80$. Consequently, the results suggest that performing multiple comparisons with RT data is less than ideal when the data distribution is characterised by a strong exponential component. In light of the good performance with distributions in which $\tau = 0$, operating an RT decomposition technique such as put forward by Heathcote (1996) is strongly recommended. In fact, after performing the RT decomposition, the different adjustment methods appeared adequately powerful even with small sample sizes.

References

Andrews, S., & Heathcote, A. (2001). Distinguishing common and task-specific processes in word identification: A matter of some moment? *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27, 514-544.

Armstrong, I. T., & Munoz, D. P. (2003). Inhibitory control of eye movements during oculomotor countermanding in adults with attention-deficit hyperactivity disorder. *Experimental Brain Research*, 152, 444-452.

Balota, D. A., & Spieler, D. H. (1999). Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency. *Journal of Experimental Psychology: General*, 128, 32-55.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57, 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the False Discovery Rate in multiple testing under dependency. *Annals of Statistics*, 29, 1165-1188.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8, 3-62.

Burbeck, S. L., & Luce, R. D. (1982). Evidence from auditory simple reaction times for both change and level detectors. *Perception and Psychophysics*, 32, 117-133.

Campbell, J. I. D., & Penner-Wilger, M. (2006). Calculation latency : The μ of memory and the τ of transformation. *Memory and Cognition*, 34, 217-226.

Dell'Acqua, R., Job, R., & Grainger, J. (2001). Is global shape sufficient for automatic object identification? *Visual Cognition*, 8, 801-821.

Dudoit, S., Popper Shaffer, J., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 1, 71-103.

Ge, Y., Dudoit, S., & Speed, T. (2003). Resampling-based multiple testing for microarray data analysis. Technical Report \#633 of UCB Stat}. <http://www.stat.berkeley.edu/~gyc>.

Heathcote, A. (1996). RTSYS: A DOS application for the analysis of reaction time data. *Behavior Research Methods, Instruments and Computers*, 28, 427-445.

- Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychological Bulletin*, 109, 340-347.
- Hohle, R.H. (1965). Inferred components of reaction times as functions of foreperiod duration. *Journal of Experimental Psychology*, 69, 382-386.
- Horn, M., & Dunnett, C. W. (2004). Power and sample size comparisons of stepwise FWE and FDR controlling test procedures in the normal many-one case. In Y. Benjamini, F. Bretz, & S. Sarkar (Eds.), *Recent developments in multiple comparison procedures*. Lecture Notes – Monograph Series. Vol. 47 (pp. 48-64). Beachwood, OH: Institute of Mathematical Statistics.
- Jöreskog, K. G., & Sörbom, D. (1996). *PRELIS 2: User's reference guide*. Scientific Software, Chicago, IL.
- Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to Familywise and Comparisonwise Type I error control. *Psychological Methods*, 4, 58-69.
- Klein, R. M. (2000). Inhibition of return. *Trends in Cognitive Sciences*, 4, 138-147.
- Leth-Steensen, C., Elbaz, Z. K., & Douglas, V. I. (2000). Mean response times, variability, and skew in the responding of ADHD children: a response time distributional approach. *Acta Psychologica*, 104, 167-190.
- Lu, C.-H., & Proctor, R. W. (1995). The influence of irrelevant location information on performance: A review of the Simon and spatial Stroop effects. *Psychonomic Bulletin and Review*, 2, 174-207.
- Madden, D. J., Gottlob, L. R., Denny, L. L., Turkington, T. G., Provenzale, J. M., Hawk, T. C. & Coleman, R. E. (1999). Aging and recognition memory: Changes in regional cerebral blood flow associated with components of reaction time distributions. *Journal of Cognitive Neuroscience*, 11, 511-520.
- Matsumoto, M., & Nishimura, T. (1998). Mersenne Twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation*, 8, 3-30.
- McGill, W. J. (1963). Stochastic latency mechanisms. In R. D. Luce, R. R. Busch, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 309-360). New York: John Wiley and Sons.
- Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental Psychology: General*, 129, 262-284.
- Mewhort, D. J. K., Braun, J. G., & Heathcote, A. (1992). Response time distributions and the Stroop task: A test of the Cohen, Dunbar, and McClelland (1990) model. *Journal of Experimental Psychology: Human Perception and Performance*, 18, 872-882.
- Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading: Visual word recognition* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., (p. 264-336).
- Penner-Wilger, M., Leth-Steensen, C., & LeFevre, J.-A. (2002). Decomposing the problem-size effect: A comparison of response time distributions across cultures. *Memory and Cognition*, 30, 1160-1167.
- R Development Core Team (2003). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, 85, 59-108.
- Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological Bulletin*, 86, 446-461.
- Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114, 510-532.
- Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, 83, 190-214.
- Ramsey, P. H. (1978). Power differences between pairwise multiple comparisons. *Journal of the American Statistical Association*, 363, 479-485.
- Reiner, A., Yekutieli, D., & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics*, 19, 368-375.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. *Behavior Research Methods, Instruments and Computers*, 33, 457-469.

Spieler, D. H., Balota, D. A., & Faust, M. E. (2000). Levels of selective attention revealed through analyses of response time distributions. *Journal of Experimental Psychology: Human Perception and Performance*, 26, 506-526.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of Benjamini-Hochberg procedure for controlling the False Discovery Rate in multiple comparison. *Journal of Educational and Behavioral Statistics*, 27, 77-83.

Van Selst, M., & Jolicoeur, P. (1994). A solution to the effect of sample size on outlier elimination. *Quarterly Journal of Experimental Psychology*, 47A, 631-650.

Van Zandt, T. (2000). How to fit a response time distribution. *Psychonomic Bulletin and Review*, 7, 424-465.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted (Vol. Ed.) & H. Pashler (Series Ed.) *Stevens' Handbook of Experimental Psychology* (3rd Edition), Volume 4: Methodology in Experimental Psychology (pp. 461-516). New York: Wiley Press.

West, R., Murphy, K. J., Armilio, M. L., Craik, F. I. M., & Stuss, D. T. (2002). Lapses of intention and performance variability reveal age-related increases in fluctuations of executive control. *Brain and Cognition*, 49, 402-419.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing*, Wiley, NY.

Wilcox, R. (2003). *Applying contemporary statistical techniques*. San Diego, Academic Press.

Wilcox, R. (2005). *Introduction to robust estimation and hypothesis testing*. Academic Press, MA.

Yekutieli, D., & Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82, 171-196.

Logit Estimation Using Warner's Randomized Response Model

Zawar Hussain Javid Shabbir
Quaid-i-Azam University

A modified hidden logit estimation procedure is presented based on Warner (1965) randomized response model. Monte Carlo simulations explore the behavior of this estimator and compare its performance with the ordinary logits estimator. Warner's model is more protective and less jeopardizing.

Key words: Logit estimation, randomized response, sensitive character.

Introduction

Binary data have been used quite frequently in econometric modeling. In the early days of econometrics these data were on the explanatory variables named as dummy variables. The development of linear and nonlinear econometrics, now, provided the ways to analyze the discrete dependent variables in regression models. They lead to the probit model and logit model. One of the assumptions in these procedures is that the empirical observations on dichotomous dependent variables are real reflections of the true values of the dependent variable. This is somewhat unrealistic assumption when modeling self-reported data on sensitive topics, such as when survey respondents are asked about embarrassing behavior, or illegal activities. Innocuous questions receive higher response rates than

questions on sensitive items, particularly on those involving perceived stigmatizing matters.

The latter often results in either refusal to respond or falsified answers. Due to this, non-response error is introduced and results in the unreliable estimation of population parameters of the interest. The reason of falsification of answer or refusal to answer might be the incentives for the survey respondents in the form of not getting embarrassed or not to be stigmatized. Corstange (2004) noted, "If the problem is that people have incentives to hide their true opinions or behavior from the interviewer, then our science suffers unless we can develop means to nullify these incentives. Survey respondents may not be willing to reveal their true answers to sensitive questions without foolproof guarantees of anonymity – not only from outside observers such as law enforcement or friends and family, but even from the interviewers themselves" (p. 5).

To nullify these adverse incentives, Corstange (2004) discussed changing the wordings of the sensitive question. But changing the statement of the question is actually changing the question and revised statements may not fully deliver the true underlying concept we hope to measure. As a means of guaranteeing anonymity to the respondent, consider Warner's (1965) randomized response model.

The randomized response models originated with Warner (1965), a statistician by discipline, and have since been improved upon by various others. Corstange (2004) stated that surprisingly enough, the procedure was almost entirely unknown among political scientists: other than a few brief research notes published

Zawar Hussain, Ph. D., is a Lecturer in the Department of Statistics, Quaid-i-Azam. His research interests include randomized response models, unequal probability sampling, Regression analysis, Bayesian estimation, Item count techniques. Email: zhlangah@yahoo.com. Javid Shabbir is Associate Professor at the University of Kent, UK, and Head of Department of Statistics, Quaid-i-Azam University. His research interests include randomized response models, unequal probability sampling, Sampling techniques. Email: jsqau@yahoo.com.

Figure 1. Graphical representation of Corstange (2004) model.

in the late 1970s, randomized response remains relatively terra incognita to the discipline. The reason of this unpopularity of randomized response among the psychologists and politicians might be that, formerly, at best they could estimate population means rather than explanatory models. In other words, they were only able to estimate the proportion of respondents who evaded taxes in the last year without being able to estimate the effects of other characteristics such as family size, race, and number of earning hands, locality, and socio-economic status on tax evasion.

Corstange's (2004) Hidden Logits

The randomized response model used by Corstange (2004) is as follows: Consider the following procedure to a yes/no question where “yes” the sensitive answer is: the respondent flips a coin and does not reveal the result to the interviewer. If the coin comes up *heads*, the respondent answers “yes” unconditionally, but if the coin comes up *tails*, the respondent answers the given yes/no question. Under these conditions, the interviewer does not know – and will never know – whether a “yes” response came as a result of a *heads* or as an answer to the question being asked. Generally, if φ is the probability of an unconditional “yes” response (in the example, $\varphi = .50$, the probability of

getting *heads*) and $(1-\varphi)$ is the probability of an actual *answer* (either “yes” or “no”), then we can represent the extensive form of the possible outcomes as in Figure 1.

From the above displayed data generating process

$$prob(yes) = \hat{\pi} = \varphi + (1-\varphi).\pi \quad (1)$$

On simplification,

$$\pi = \frac{\hat{\pi} - \varphi}{1 - \varphi} \quad (2)$$

In ordinary logit models

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i\boldsymbol{\beta} \quad (3)$$

where $\mathbf{x}_i$ is the row vector of observations on explanatory variables and $\boldsymbol{\beta}$ is the column vector of parameters. From equation (3) we can observe that estimation of the $\boldsymbol{\beta}$'s is not possible because there are not any data on π_i . The only data available are on the explanatory variables and $\hat{\pi}_i$. Therefore, in order to move further express the logit model in terms of information available (i.e., $\hat{\pi}_i$).

LOGIT ESTIMATION

Figure 2 : Graphical representation of Warner's(1965) RRM.

Substituting (2) in (3) and solving for π

$$\hat{\pi}_i = \frac{e^{x_i\beta} + \varphi}{1 - \varphi} \quad (4)$$

Using equation (4) it is possible to estimate our parameters of interest, β 's, and thereafter the logits by maximum likelihood method. As shown by Corstange (2004), setting the derivatives of the likelihood function equal to zero maximizes the likelihood function but that equation cannot be solved analytically. Therefore, it is solved numerically.

Deviation of Modified Hidden Logit

The Warner's (1965) randomized response model provides more privacy and anonymity to the respondents than provided by the randomized response model used by Corstange (shown below). Our Modified hidden logits are based on Warner's (1965) randomized response model. Warner's (1965) randomized response device consists of two complementary statements, say, A and A^c . The statements A and A^c are presented with probabilities P and $1-P$ respectively. The respondents are required to select one of the two statements randomly and answer yes or no according to their true status. The extensive form of the outcomes of Warner's device is shown in Fig.2.

The probability of a yes answer in Warner's (1965) device is

$$\begin{aligned} \text{Prob}(\text{yes}) = \\ P.\pi + (1-P).(1-\pi) = \theta \quad (\text{say}). \end{aligned} \quad (5)$$

Then using the steps of equations (2) and (3)

$$\pi = \frac{\theta - (1-2P)}{(2P-1)} \quad (6)$$

On substituting equation(6) in equation(3)

$$\theta_i = \frac{P.e^{x_i\beta} + (1-P)}{1 + e^{x_i\beta}} \quad (7)$$

For $P=1.0$ it becomes the ordinary logits derived from direct response.

Because of the interest in the estimation of β 's, and all the information available is the observed probability of "yes" response, θ_i , the estimation is conducted using θ_i .

Suppose y_i is a binary random variable taking two values, '0'(no) and '1'(yes) with probabilities $1-\theta_i$ and θ_i respectively, then

Table 1.(a)

$N = 1000$	$\tilde{\beta}_i$	Ordinary Logit $P = 1.0$	Modified hidden Logit. $P = 0.10$	Modified hidden Logit. $P = 0.20$	Modified hidden Logit. $P = 0.25$	Modified hidden Logit. $P = 0.40$
x_0	0.0	0.00046 (0.1071)*	-0.004 (0.138)	0.006 (0.193)	0.0078 (0.244)	0.0056 (0.245)
x_1	1.0	1.014 (0.0914)	1.018 (0.132)	1.040 (0.201)	1.075 (0.274)	1.081 (0.3001)
x_2	1.0	1.014 (0.093)	1.019 (0.129)	1.037 (0.2009)	1.0706 (0.272)	1.082 (0.299)
x_3	1.0	1.012 (0.093)	1.018 (0.1302)	1.038 (0.2013)	1.0182 (0.279)	1.034 (0.2987)

Table 1. (b)

$N=2000$	$\tilde{\beta}_i$	Ordinary Logit $P = 1.0$	Modified hidden Logit. $P = 0.10$	Modified hidden Logit. $P = 0.20$	Modified hidden Logit. $P = 0.30$	Modified hidden Logit. $P = 0.40$
X_0	0.00	-0.0001 (0.070)	-0.001 (0.090)	0.0003 (0.125)	0.0015 (0.200)	0.0016 (0.211)
X_1	1.00	1.008 (0.064)	1.011 (0.092)	1.019 (0.136)	1.051 (0.231)	1.055 (0.223)
X_2	1.00	1.006 (0.064)	1.010 (0.092)	1.018 (0.135)	1.051 (0.228)	1.054 (0.311)
X_3	1.00	1.006 (0.063)	1.011 (0.091)	1.019 (0.136)	1.051 (0.233)	1.049 (0.291)

Table 1. (c)

$N=5000$	$\tilde{\beta}_i$	Ordinary Logit $P = 1.0$	Modified hidden Logit. $P = 0.10$	Modified hidden Logit. $P = 0.20$	Modified hidden Logit. $P = 0.30$	Modified hidden Logit. $P = 0.40$
X_0	0.00	0.0006 (0.046)	0.0004 (0.059)	0.00004 (0.081)	-0.001 (0.125)	0.0016 (0.192)
X_1	1.00	1.0010 (0.040)	1.001 (0.057)	1.005 (0.082)	1.016 (0.131)	1.025 (0.183)
X_2	1.00	1.002 (0.039)	1.001 (0.056)	1.005 (0.082)	1.017 (0.132)	1.024 (0.194)
X_3	1.00	1.002 (0.040)	1.002 (0.056)	1.007 (0.080)	1.019 (0.132)	1.029 (0.165)

LOGIT ESTIMATION

Table 1. (d)

$N=10000$	$\tilde{\beta}_i$	Ordinary Logit $P=1.0$	Modified hidden Logit. $P=0.10$	Modified hidden Logit. $P=0.20$	Modified hidden Logit. $P=0.30$	Modified hidden Logit. $P=0.40$
X_0	0.00	0.0001 (0.031)	0.0019 (0.042)	0.0013 (0.058)	0.0015 (0.089)	-0.0081 (0.200)
X_1	1.00	1.001 (0.028)	1.001 (0.040)	1.002 (0.057)	1.006 (0.092)	1.061 (0.212)
X_2	1.00	1.001 (0.028)	1.002 (0.038)	1.004 (0.056)	1.008 (0.091)	1.060 (0.199)
X_3	1.00	1.0004 (0.028)	1.0009 (0.040)	1.001 (0.055)	1.004 (0.090)	1.071 (0.187)

given y_i , the likelihood function of β is given by

$$L(\beta/y_i) = \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i} \quad (8)$$

and by taking natural logarithm on both sides

$$\ell = \ln L(\beta/y_i) = \sum_{i=1}^n \{y_i \cdot \ln \theta_i + (1 - y_i) \cdot \ln(1 - \theta_i)\} \quad (9)$$

The first order derivative of above equation with respect to the parameter vector β is given by

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i (P-1) \left(\frac{(3-2P+e^{x_i \tilde{\beta}})}{(1-P+Pe^{x_i \tilde{\beta}})(2-P+(1-P)e^{x_i \tilde{\beta}})(1+e^{-x_i \tilde{\beta}})} \right) - \frac{(P-1)}{((1-P)e^{x_i \tilde{\beta}}+2-P)(1+e^{-x_i \tilde{\beta}})} \right\} x_i \quad (10)$$

When this equation is set equal to zero it maximizes the log-likelihood function but this equation cannot be solved analytically (see appendix). Therefore, its numerical solution may be obtained.

Comparison Of Modified Hidden Logits With Ordinary Logits

For comparison purposes, a small sample simulation study was conducted and results are given in table 1(a). The reason for small sample study is that the properties of consistency, normality and efficiency are well established for all maximum likelihood estimators (Green, 2000, & King, 1998). However, to see the pattern in the variances of β_i results for $N=2000, 5000$, and 10000 are presented in Table 1(b,c,d). The data presented here were generated as follows. For each $P, 1000, 2000, 5000$ and 10000 samples were generated from a three regressors equation with no constant term. For simplicity, each $\beta_i = 1$. Also each $x_i \sim U(-3, 3)$.

Given the above experimental conditions modified hidden logit return $\hat{b}_i$ that quite closely track the true population parameters β_i . The Table 1 (a, b, c, d) also compare the performance of the modified hidden logit estimator with ordinary logit (when $P=1.0$) at selected levels of P . From the Table 1(a, b, c, d) it is clear that modified hidden logit quite closely track the true β 's but at the cost of increased variances.

LOGIT ESTIMATION

Figure 3. Graphs of $\hat{\beta}_i$'s against P for $N = 1000, 2000, 5000$ and 10000 .

Behavior of β_i 's with P

From Figure 3 it is apparent that the modified hidden logit estimates of β_i 's deviate upward from the true β_i 's as p moves from 0.0 to 0.5 and then become close to the true parameters β_i 's as p increases from 0.5 to 1.0. An important point to remember is that when $P = 0.5$ the estimators of β_i 's do not exist (as is the case of applying Warner's model to estimate π , the proportion of population with sensitive attribute). It is interesting to note that standard errors of the β_i 's are symmetric around $P = 0.5$. When the values of P moves away from 0.5 the standard errors of all the β_i 's decreases.

Respondent's Protection

Three basic concerns in randomized response models are jeopardy, suspicion, and efficiency. Jeopardy is the extent to which an affirmative answer implies the sensitive attribute; that is, the likelihood that the person has the attribute, given a yes response. In forced alternatives (answer either the sensitive or non-sensitive question), jeopardy increases as the probability that sensitive question was asked increases and the percentage of the population with the sensitive character decreases.

Suspicion is the extent to which a negative answer implies the sensitive attribute; that is, the likelihood that a person has the attribute, given a response. In forced alternatives (answer either the sensitive or non-sensitive question), suspicion increases as the probability that the sensitive question was asked decreases and the percentage of population with the non-sensitive character also decreases.

Efficiency is the loss in precision as a result of randomized response technique. It increases as the probability that the sensitive question was asked decreases.

In comparing the randomized response models emphasis has been on the variances. Greenberg, Abul-El, Simmons, and Horvitz(1969), Moors(1971), and Dowling and Shachtman(1975) are some of many to be referred. The emphasis on variances amounts to

considering the matters from statistician's point of view only. Whereas the respondent's interest would be in the extent to which the different methods provide protection against their privacy. Leysieffer and Warner (1976), and Lanke (1975,76) provided the measures of protection provided by the different methods. Leysieffer and Warner (1976) proposed the natural measure of Jeopardy carried by a response R (either yes = Y or no = N), about A and A^c respectively, which are as $g(R/A) = P(R/A)/P(R/A^c)$ and $g(R/A^c) = 1/g(R/A)$, where A and A^c are defined as above. These functions are called jeopardy functions. And the particular response R is jeopardizing if $g(R/A) = 1$.

Lanke(1976) proposed a measure of suspicion defined as

$$\psi = \max(P(A/Y), P(A/N)), \text{ where } P(A/Y) \text{ and } P(A/N)$$

are conditional probabilities of belonging to a sensitive group A given a particular response Y or N , and proposed that a method is more protective for which

$$\psi = \max(P(A/Y), P(A/N))$$

is smaller.

These two measures are calculated for both of the randomized response models used by Corstange(2004), and Warner(1965) which are as follows:

(i) For Warner's model

$$g_w(Y/A) = \frac{P}{1-P}$$

and

$$\psi_w = \max(P(A/Y), P(A/N))$$

(ii) For Corstange model

$$g_d(Y/A) = \infty$$

and

$$\psi_d = \max(P(A/Y), P(A/N)) = 1.$$

It can be seen that $g_w(Y/A) \leq g_d(Y/A)$ and $\psi_w \leq \psi_d$. It suggests that Warner's model is less jeopardizing and more protective.

LOGIT ESTIMATION

Figure 4. The behavior of standard errors of the $\hat{b}_i$'s with increasing values of P for $N = 1000$.

Choice of p

Setting a desirable value of P depends upon the nature of population. As we have just discussed above that there are three major concerns of using randomized response techniques: jeopardy, suspicion and efficiency. Jeopardy increases with the increase in P and decrease in the proportion of population possessing sensitive character whereas suspicion increases with the decrease in P and the increase in the proportion of population possessing sensitive character. It has been showed that $g_w(Y/A) \leq g_d(Y/A)$ and $\psi_w \leq \psi_d$, so Warner's randomizing device is superior to that of Corstange's. Form table 2 it is apparent that the larger standard errors of each β_i when P is closer to 0.5. Thus a value farther from 0.5 should be set which seems desirable (e, g. 0.3, 0.4, 0.6 or 0.7) and creates a balance between jeopardy and suspicion. In connection with isolated study of comparing the hidden logits based on Corstange (2004)'s model and

Warner's (1965) model we suggest to set P at smaller level as it would provide more anonymity and would be less jeopardizing. As far as suspicion is concerned, Warner's (1965) model induces less suspicion for every P . So Warner's model would be a better choice as compared to Corstange (2004) model.

Fig. 4 presents the behavior of standard errors of the estimators $\hat{b}_i, i = 1, 2, 3$. for different values of P . It can be easily seen that when P is closer to 0.5 the standard errors of the estimates are larger and setting P closer to 0.5 would induce unreliability in the estimates. Therefore, we suggest setting P away from 0.5. The same behavior of standard errors with respect to changes in P is observed for other values of N .

Discussion

As survey statisticians, our interest in sensitive topics inevitably leads us to ask sensitive questions. As this article shows, however, we must take care when we study such topics,

especially when drawing inferences from self-reported, falsifiable answers to questions. By falsifying the true responses respondents get incentives by misrepresenting them. Randomized response as a questioning technique allows us, at least in principle, to nullify these incentives. The estimator developed here allows us to model questions of this nature, and simulations suggest that proceeding in this fashion allows us to draw more valid and more useful inferences about sensitive social issues.

References

Corstange, D. (2004). Sensitive questions, truthful responses? Randomized response and hidden logit as a procedure to estimate it." 2004 Annual Meeting of the American Political Science Association September 2 - September 5, 2004. Available at <http://www.umich.edu/~dancorst>.

Greene, W. (2000/1993). *Econometric Analysis*. (4th ed.). Upper Saddle River, NJ: Prentice Hall.

King G. (1998/1989). *Unifying political methodology*. Ann Arbor: University of Michigan Press.

Warner, S. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, 60(309), 63-69.

Appendix.

Derivation of equation (4).

For

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = x_i\beta$$

put the value of π_i from Warner's model and get:

$$\frac{\left(\frac{\theta_i - (1-P)}{(2P-1)}\right)}{\left(\frac{P-\theta_i}{(2P-1)}\right)} = e^{x_i\beta}$$

$$\frac{(\theta_i - (1-P))}{(P - \theta_i)} = e^{x_i\beta}$$

$$\theta_i - (1-P) = Pe^{x_i\beta} - \theta_i e^{x_i\beta}$$

$$\theta_i = \frac{Pe^{x_i\beta} + (1-P)}{(1 + e^{x_i\beta})}$$

first derivative of θ_i with respect to true $\tilde{\beta}$

$$\frac{\partial \theta_i}{\partial \tilde{\beta}} = \frac{\partial \left(\frac{Pe^{x_i\tilde{\beta}} + (1-P)}{1 + e^{x_i\tilde{\beta}}} \right)}{\partial \tilde{\beta}}$$

$$\frac{\partial \hat{\pi}}{\partial \tilde{\beta}} = \frac{\partial \left(\frac{Pe^{x_i\tilde{\beta}}}{1 + e^{x_i\tilde{\beta}}} \right)}{\partial \tilde{\beta}} + \frac{\partial \left(\frac{1-P}{1 + e^{x_i\tilde{\beta}}} \right)}{\partial \tilde{\beta}}$$

$$= \frac{(1 + e^{x_i\tilde{\beta}})Px_i e^{x_i\tilde{\beta}} - Px_i e^{2x_i\tilde{\beta}}}{(1 + e^{x_i\tilde{\beta}})^2} + \frac{-x_i e^{x_i\tilde{\beta}}}{(1 + e^{x_i\tilde{\beta}})^2}$$

$$= \frac{Px_i e^{x_i\tilde{\beta}} - x_i e^{x_i\tilde{\beta}}}{(1 + e^{x_i\tilde{\beta}})^2}$$

$$\frac{\partial \theta_i}{\partial \tilde{\beta}} = \frac{(P-1)x_i e^{x_i\tilde{\beta}}}{(1 + e^{x_i\tilde{\beta}})^2}$$

Likelihood function : observed θ_i with respect to true $\tilde{\beta}$.

$$L(\beta / y_i) = \prod_{i=1}^n \theta_i^{y_i} (1 - \theta_i)^{1-y_i}$$

$$\ell = \ln L(\tilde{\beta} / y_i) = \sum_{i=1}^n \{y_i \cdot \ln \theta_i + (1 - y_i) \cdot \ln(1 - \theta_i)\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i \theta_i^{-1} \left(\frac{\partial \theta_i}{\partial \tilde{\beta}} \right) + (1 - y_i) (1 - \theta_i)^{-1} \left(\frac{\partial (1 - \theta_i)}{\partial \tilde{\beta}} \right) \right\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i \left(\frac{1+e^{x_i \tilde{\beta}}}{1-P+Pe^{x_i \tilde{\beta}}} \right) \left(\frac{(P-1)x_i e^{x_i \tilde{\beta}}}{(1+e^{x_i \tilde{\beta}})^2} \right) - (1-y_i) \left(\frac{1+e^{x_i \tilde{\beta}}}{(1-P)e^{x_i \tilde{\beta}}+2-P} \right) \left(\frac{(P-1)x_i e^{x_i \tilde{\beta}}}{(1+e^{x_i \tilde{\beta}})^2} \right) \right\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i \left(\frac{(P-1)x_i e^{x_i \tilde{\beta}}}{(1-P+Pe^{x_i \tilde{\beta}})(1+e^{x_i \tilde{\beta}})} \right) - (1-y_i) \left(\frac{(P-1)x_i e^{x_i \tilde{\beta}}}{((1-P)e^{x_i \tilde{\beta}}+2-P)(1+e^{x_i \tilde{\beta}})} \right) \right\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i x_i \left(\frac{(P-1)}{(1-P+Pe^{x_i \tilde{\beta}})(1+e^{-x_i \tilde{\beta}})} \right) + y_i x_i \frac{(P-1)}{((1-P)e^{x_i \tilde{\beta}}+2-P)(1+e^{-x_i \tilde{\beta}})} - \frac{(P-1)}{((1-P)e^{x_i \tilde{\beta}}+2-P)(1+e^{-x_i \tilde{\beta}})} \right\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i x_i (P-1) \left(\frac{(3-2P+e^{x_i \tilde{\beta}})}{(1-P+Pe^{x_i \tilde{\beta}})(2-P+(1-P)e^{x_i \tilde{\beta}})(1+e^{-x_i \tilde{\beta}})} \right) - \frac{(P-1)}{((1-P)e^{x_i \tilde{\beta}}+2-P)} x_i (1+e^{-x_i \tilde{\beta}})^{-1} \right\}$$

$$\frac{\partial \ell}{\partial \tilde{\beta}} = \sum_{i=1}^n \left\{ y_i (P-1) \left(\frac{(3-2P+e^{x_i \tilde{\beta}})}{(1-P+Pe^{x_i \tilde{\beta}})(2-P+(1-P)e^{x_i \tilde{\beta}})(1+e^{-x_i \tilde{\beta}})} \right) - \frac{(P-1)}{((1-P)e^{x_i \tilde{\beta}}+2-P)} (1+e^{-x_i \tilde{\beta}})^{-1} \right\} x_i$$

Estimating How Many Observations are Needed to Obtain a Required Level of Reliability

David A. Walker
Northern Illinois University

This article provides a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. A SPSS program is provided for users to analyze situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

Key words: Spearman-Brown, reliability, observations.

Introduction

The Spearman-Brown Prophecy Formula (SBPF) is often employed to estimate split-half reliability, as a function of internal consistency, of the variability of scores split on a composite test, and based on the assumptions that the two halves of a test have equal variance parameters and are consistent in content (Kristof, 1974; Zimmerman, 1970). An application of the Spearman-Brown formula is to estimate how many items need to be added to a test to obtain a specified level of reliability (Burnett, 1974). Li and Wainer (1997) noted that the Spearman-Brown formula's principal use has been to obtain, "... the reliability coefficient for a composite measurement as the sum of n individual measurements..." (p. 479). Its calculation is used as a function of estimating the score reliability of lengthened or shortened tests. The general formula for the SBPF, given by Krathwohl (1993), is expressed as

$$r = (kr) / [1 + (k - 1)r] \quad (1)$$

David Walker is an Associate Professor of Educational Research. His Interests include structural equation modeling, effect sizes, factor analysis, predictive discriminant analysis, and bootstrapping. Email him at dawalker@niu.edu.

where, k = the ratio of items in the new test to those in the original form; r = the average of the sample correlations among individual measures. A simplified version of the Spearman-Brown Prophecy Formula, noted by Charter (2001) can be expressed as:

$$r_{kk} = 2r_{12} / (1 + r_{12}) \quad (2)$$

where, r_{xy} = the correlation between the two halves of a test;

As an extension of the use of the Spearman-Brown formula, classroom observations and raters' judgments have been added to this application of it by expanding its use to situations for estimating the reliability of pooled judgments or observations (cf. Blok, 1985; Jenkins, Bausell, & Magoon, 1972). Hartmann (1976) describes these instances as $N = 1$ designs, which are "specifically relevant to reliability assessment [and] involve sessions, observers, and trails (multiple brief observation periods) within sessions" (p. 844).

There are various sources of error affiliated with classroom observational data or pooled judgments. For example, but not all inclusive, error can be derived from the length of an observation, with shorter observations a prevailing source of error; from a lack of equivalence between raters, which is often difficult to obtain with consecutive observational tasks; from observational processes that may cause variability among raters; from inter-rater disagreement; or from large deviations in performers' performances across observational points (Blok, 1985; Hartmann, 1976; McGaw,

Wardrop, & Bunda, 1972; Rogosa & Ghandour, 1991; Rowley, 1978).

To minimize sources of error, Rowley (1978) found that the pooling together of observational periods so that they occur more frequently, instead of prolonged observations, is more beneficial to reliability, "Reliability will be enhanced by a more representative sampling of occasions, and this is best achieved by using a larger number of shorter observation periods" (p. 172). Medley and Mitzel (1963) determined that an increase in congruent observational periods, but not an increase in observers, could lessen measurement error. Finally, Meehl (1999) found that judges' ratings pertaining to a common objective, or pooling their judgments, can increase reliability and is a beneficial technique: "If we have the judgments of only a few scientists (rating a batch of theories of single experiment), we can estimate the reliability of a larger pooled judgment via the Spearman-Brown Prophecy Formula.... to predict the boosted reliability of a lengthened mental test, [it] has turned out to be quite accurate when the elements are not test items but human judgments." (p. 292)

Purpose

The purpose of this study is to provide researchers with a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. As well, SPSS (Statistical Package for the Social Sciences) syntax is provided in Appendix A for users to create Table 1 or analyze other situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

Results

As can be seen from Table 1, when the initial reliability from 1 observation is very low, ranging from .100 to .200, it would require between 81 (i.e., $r = .100$) to 36 (i.e., $r = .200$) observations to increase reliability to a level of .900, respectively. Further review of Table 1

indicates that as the initial reliability measure increases into the moderate range (e.g., $\geq .600$), the number of observations needed to enhance reliability would decrease, which is to be expected. Further, the data in Table 1 can be used as a scale by researchers involved in observational types of studies to determine, based on a preliminary measure of reliability, how many more observations would be required to reach a required level of reliability.

Usage Example

Assuming that many of the potential sources of error noted previously with use of this form of the Spearman-Brown formula were addressed and the user understood the tenets of reliability in terms of employment of observational protocols, calculation, and the interpretation of results; and the contextual uses of the SBPF versus coefficient alpha, for instance, for certain applications (cf. Charter, 2001; Martin, 1977), which admittedly may not be the case in every situation, the use of the Spearman-Brown formula to estimate the reliability of pooled judgments or observations may be warranted. For example, if a college or university-level researcher were conducting classroom-based research and performed an initial observation in a class that lasted for 20 minutes and obtained a score reliability estimate of .600 derived from the protocol used in the observation (i.e., the left column of Table 1), to increase the reliability to a desired level of .800 (i.e., the center column of Table 1), the researcher would need 3 more congruent observational periods (i.e., the right column of Table 1).

Rowley (1978) demonstrated this concept in a much more truncated example than Table 1, where it was determined that "... we may observe that a reliability of .176 obtained from one 10-minute visit could be increased to516 by making five times as many visits" (p. 170). Rowley's example can be replicated in the syntax in Appendix A by entering in the initial reliability level of .176 in the left column between the BEGIN DATA and END DATA field, putting in the desired reliability level of .516 in the right column of the same field, and then running the program, which will produce the number of observational periods needed of 5.

OBSERVATIONS NEEDED TO OBTAIN A LEVEL OF RELIABILITY

Table 1. The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed	Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.1	0.2	2	0.2	0.3	2
0.1	0.25	3	0.2	0.35	2
0.1	0.3	4	0.2	0.4	3
0.1	0.35	5	0.2	0.45	3
0.1	0.4	6	0.2	0.5	4
0.1	0.45	7	0.2	0.55	5
0.1	0.5	9	0.2	0.6	6
0.1	0.55	11	0.2	0.65	7
0.1	0.6	13	0.2	0.7	9
0.1	0.65	17	0.2	0.75	12
0.1	0.7	21	0.2	0.8	16
0.1	0.75	27	0.2	0.85	23
0.1	0.8	36	0.2	0.9	36
0.1	0.85	51	0.2	0.95	76
0.1	0.9	81	0.25	0.35	2
0.1	0.95	171	0.25	0.4	2
0.15	0.25	2	0.25	0.45	2
0.15	0.3	2	0.25	0.5	3
0.15	0.35	3	0.25	0.55	4
0.15	0.4	4	0.25	0.6	4
0.15	0.45	5	0.25	0.65	6
0.15	0.5	6	0.25	0.7	7
0.15	0.55	7	0.25	0.75	9
0.15	0.6	9	0.25	0.8	12
0.15	0.65	11	0.25	0.85	17
0.15	0.7	13	0.25	0.9	27
0.15	0.75	17	0.25	0.95	57
0.15	0.8	23	0.3	0.4	2
0.15	0.85	32	0.3	0.45	2
0.15	0.9	51	0.3	0.5	2
0.15	0.95	108	0.3	0.55	3

Table 1 (cont'). The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed	Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.3	0.6	4	0.45	0.8	5
0.3	0.65	4	0.45	0.85	7
0.3	0.7	5	0.45	0.9	11
0.3	0.75	7	0.45	0.95	23
0.3	0.8	9	0.5	0.6	1
0.3	0.85	13	0.5	0.65	2
0.3	0.9	21	0.5	0.7	2
0.3	0.95	44	0.5	0.75	3
0.35	0.45	2	0.5	0.8	4
0.35	0.5	2	0.5	0.85	6
0.35	0.55	2	0.5	0.9	9
0.35	0.6	3	0.5	0.95	19
0.35	0.65	3	0.55	0.65	2
0.35	0.7	4	0.55	0.7	2
0.35	0.75	6	0.55	0.75	2
0.35	0.8	7	0.55	0.8	3
0.35	0.85	11	0.55	0.85	5
0.35	0.9	17	0.55	0.9	7
0.35	0.95	35	0.55	0.95	16
0.4	0.5	1	0.6	0.7	2
0.4	0.55	2	0.6	0.75	2
0.4	0.6	2	0.6	0.8	3
0.4	0.65	3	0.6	0.85	4
0.4	0.7	3	0.6	0.9	6
0.4	0.75	4	0.6	0.95	13
0.4	0.8	6	0.65	0.75	2
0.4	0.85	8	0.65	0.8	2
0.4	0.9	14	0.65	0.85	3
0.4	0.95	28	0.65	0.9	5
0.45	0.55	1	0.65	0.95	10
0.45	0.6	2	0.7	0.8	2
0.45	0.65	2	0.7	0.85	2
0.45	0.7	3	0.7	0.9	4
0.45	0.75	4	0.7	0.95	8

OBSERVATIONS NEEDED TO OBTAIN A LEVEL OF RELIABILITY

Table 1 (cont'). The Number of Observations Needed to Obtain an Increased Reliability Coefficient

Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed	Initial Reliability Estimate	Increased Reliability Estimate	Observations Needed
0.75	0.85	2	0.8	0.95	5
0.75	0.9	3	0.85	0.9	2
0.75	0.95	6	0.85	0.95	3
0.8	0.9	2	0.9	0.95	2

This article provides a detailed table containing estimations of how many observations are needed to obtain an increased reliability coefficient for situations such as observational data collection in the classroom. As well, SPSS syntax is provided for users to analyze situations where an initial reliability value is obtained and the user wants to determine how many more observations are needed to reach a required level of reliability.

This article could be of use to researchers who carry-out school-based research studies, those who conduct classroom-based observations, for example, of student teachers, student engagement, leadership capacity, or those engaged in decision-making studies related to a specified criterion. Thus, the merit in the use of the program in Appendix A or Table 1 is to assist researchers with an easily understood method to determine if the initial score reliability from an observational protocol used in a classroom to measure a particular trait or performance is on target or are further, congruent observational periods needed to reach a desired level of score reliability.

References

- Blok, H. (1985). Estimating the reliability, validity, and invalidity of essay ratings. *Journal of Educational Measurement*, 22, 41-52.
- Burnett, J. D. (1974). Parallel measurements and the Spearman-Brown Formula. *Educational and Psychological Measurement*, 34, 785-788.

Charter, R. A. (2001). It is time to bury the Spearman-Brown "prophecy" formula for some common applications. *Educational and Psychological Measurement*, 61, 690-696.

Hartmann, D. P. (1976). Some restrictions in the application of the Spearman-Brown Prophecy Formula to observational data. *Educational and Psychological Measurement*, 36, 843-845.

Jenkins, J. R., Bausell, R. B., & Magoon, A. J. (1972). Selection of prose material for testing. *Journal of Educational Measurement*, 9, 97-103.

Krathwohl, D. R. (1993). *Methods of educational and social science research: An integrated approach*. White Plains, NY: Longman.

Kristof, W. (1974). Estimation of reliability and true score variance from a split of a test into three arbitrary parts. *Psychometrika*, 39, 491-499.

Li, H., & Wainer, H. (1997). Toward a coherent view of reliability in test theory. *Journal of Educational and Behavioral Statistics*, 22, 478-484.

Martin, J. (1977). The development and use of classroom observation instruments. *Canadian Journal of Education*, 2, 43-54.

McGaw, B., Wardrop, J. L., & Bunda, M. A. (1972). Classroom observation schemes: Where are the errors? *American Educational Research Journal*, 9, 13-27.

Medley, D. M., & Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), *Handbook of research on teaching* (pp. 267-269). Chicago: Rand McNally.

Meehl, P. E. (1999). How to weight scientists' probabilities is not a big problem: Comment on Barnes. *British Journal for the Philosophy of Science*, 50, 283-295.

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, 16, 157-252.

Rowley, G. L. (1978). The relationship of reliability in classroom research to the amount of observation: An extension of the Spearman-Brown formula. *Journal of Educational Measurement*, 15, 165-180.

Zimmerman, D. W. (1970). Variability of test scores and the split-half reliability coefficient. *Educational and Psychological Measurement*, 30, 259-266.

Appendix A. Syntax for Estimation of How Many Observations are Needed to Obtain an Increased Reliability Coefficient.

Author: David A. Walker (2008), dawalker@niu.edu, Northern Illinois University

DATA LIST LIST/ r REST (2F9.3).

NOTE: As the first number between BEGIN DATA and END DATA, put your initial score reliability and then as the second number, put the estimated, increased score reliability that you would like to achieve.

BEGIN DATA

.176 .516

END DATA.

COMPUTE OBS = (REST*(1-r)/(r*(1-REST))).

EXECUTE.

FORMAT OBS (F8.0).

VARIABLE LABELS r 'Single Observation Reliability'/REST 'Estimated, Boosted Reliability'/OBS 'The Number of Observations Needed to Equal an Estimated, Boosted Reliability'/.

REPORT FORMAT=LIST AUTOMATIC ALIGN (LEFT)

MARGINS (*,110)

/VARIABLES= r REST OBS

/TITLE "Estimation of How Many Observations are Needed to Obtain an Increased Reliability Coefficient".

Probability of Coverage and Interval Length for Two-Group Techniques Assessing the Median and Trimmed Mean

S. Jonathan Mends-Cole
Walden University

The purpose of the present study was to assess the probability of coverage and interval length of selected statistical techniques that have a higher finite sample breakdown point than the mean and appropriate levels of probability of coverage when using Bradley's (1978) criterion. The techniques were examined using real education and psychology datasets (Sawilowsky & Fahoome, 2003, Sawilowsky & Blair, 1992). Welch's test exhibited appropriate coverage for the smooth symmetric, mass at zero, digit preference, and extreme bimodal distributions. Yuen's technique performed well under an extreme bimodal distribution. Results concerning the Maritz-Jarrett and the McKean-Schrader techniques are also presented.

Key words: Trimmed mean, median, confidence interval, interval length, probability of coverage

Introduction

A researcher may want to know a range of values that may enclose the population parameter with a given level of confidence. A confidence interval provides a range of values that one can be $(1-\alpha)100\%$ confident the population parameter is enclosed (Sawilowsky & Fahoome, 2003, p. 200-201). The adequacy of a confidence interval is assessed through probability of coverage, $1-\hat{\alpha}$. Within education and psychology, statistical techniques have been assessed through Type I ($\hat{\alpha}$) and Type II ($\hat{\beta}$) error rates (e.g., Wilcox & Charlin, 1986, Wilcox, 1994, Luh & Guo, 2000, Wilcox, Kowalchuk, & Olejnik, 2002). Although many studies have examined techniques using error rates, some studies have examined the techniques using probability of coverage and interval length. Examples would include the studies done by Bonett and Price (2002) and Price and Bonett (2002).

Many situations motivate the use of confidence intervals. First, if no hypothesis is to be tested and one wants to know the range of plausible values for the population parameter, Knapp (1999) recommended using a confidence interval. Second, retrospective power analysis employs statistical power (a) following the statistical analysis, and (b) with a sample estimate of effect size. Statistical power is the *a priori* probability of detecting an effect if it exists, that is $\mu_1 - \mu_2 \neq 0$ (Wilcox, 1996; Zumbo & Hubley, 1998). The use of statistical power in retrospective power analysis is untenable (Zumbo & Hubley, 1998; Knapp, 1999, Hoenig & Heisey, 2001). Hoenig and Heisey (2001), and Wilkinson and Taskforce on Statistical Inference (1999) recommended using confidence intervals instead of retrospective power analysis. Confidence intervals provide an indication of statistical precision. The interval is more precise if the length is narrower (Hinkle, Wiersma, & Jurs, 1998, p. 225).

Confidence intervals also provide a test of the null hypothesis. Values outside the interval limits provide evidence for rejecting a range of null hypotheses (Hinkle, Wiersma, & Jurs, 1998, p. 224, Bonett & Price, 2002, p. 372). When comparing different statistical techniques or when applied at different alpha levels, narrower interval lengths imply greater

S. Jonathan Mends-Cole is a member of the part-time faculty at Walden University. He can be reached at sjmendscole@aim.com or at P.O. Box 07285, Detroit, MI 48207.

statistical power. Narrower interval lengths imply a smaller standard error. A smaller standard error implies that it will be more likely that the test statistic will be rejected.

Here, a study was done on the probabilities of coverage and interval length of techniques selected on the minimum number of sample values that can be altered, thus making the measure of central tendency arbitrarily large or small – the finite sample breakdown point (FSBP) (Wilcox, 2001, p. 149). An FSBP of $1/n$ is given for the mean. One outlier can make the mean arbitrarily large or small and the outlier can increase the variance. The outlier effect on the mean and variance affects the Type I and Type II error rates of statistical techniques that depend on the mean.

Consider the 0.05 alpha level. When sampling from skewed distributions where outliers occur frequently, the Type I error rate exceeds the nominal level (α) and the probability of coverage is below the nominal level ($1-\alpha$) (Wilcox, 2003). For example, Mends-cole (2006) found that probability of coverage for Students' t was less than 0.925 under inverse heteroscedasticity. The standard of 0.925 was adopted from Bradley's (1978) criterion and recommended by Bunner (2003): $1-0.5\alpha \leq 1-\hat{\alpha} \leq 1-1.5\alpha$. At the extremes of skewness ($\gamma_3 > 1.25$), probabilities of coverage for Welch's t were below 0.925. The results were similar to Algina, Oshima, and Lin (1994) and Luh and Guo (2000). Under similar conditions, the probability of coverage for Welch's technique was less than 0.925.

Preference was given to measures of central tendency with an FSBP that was higher than the mean. An FSBP of 0.20 is given for the 20% trimmed mean; the median has an FSBP of 0.50 (Wilcox, 2001, p. 149). Selected procedures for evaluating the trimmed mean and median included the confidence interval obtained by inverting Yuen's trimmed t -statistic, the Maritz-Jarrett median z -statistic (M-J), and the McKean-Schrader median z -statistic (M-S). Beyond testing a measure of central tendency with a high FSBP, the techniques have acceptable levels of Type I error rates, $0.025 \leq \hat{\alpha} \leq 0.075$. In Table 1, γ_3 denotes

skewness, γ_4 denotes kurtosis ($\gamma_4=0$ for a normal distribution), skewness and kurtosis specifications for Wilcox (1994) represent the second group data. The first group is sampled from a standard normal distribution. Skewness and kurtosis specifications for Bonett and Price (2002) represent the second group data. The first group is sampled from an exponential distribution for the first two rows and from a standard normal distribution from the third row. Under conditions of skewness, each technique exhibited minor bias in terms of Type I error rate and probability of coverage, $0.925 \leq 1-\hat{\alpha} \leq 0.975$.

Some questions arise when considering the studies presented in Table 1. (a) The techniques were recommended based on random numbers generated using mathematical functions. Results from the mathematical functions may not represent the samples observed in applied situations in education and psychology. To the extent that Monte Carlo samples represent applied situations, the results generalize to like situations (Sawilowsky & Fahoome, 2003, p. 443). (b) The techniques were recommended based on Type I and Type II error rates. The probability-coverage and interval-length are specifications of the confidence interval.

Method

Yuen's technique

Yuen's technique involves trimming and Winsorization to account for skewness. Trimming a group sample involves omitting a proportion of the largest scores and the same proportion of the smallest scores from the sample. Winsorization involves replacing a fixed proportion of the largest scores with the maximum score for the trimmed version of the same sample, and replacing an equivalent number of the smallest scores with the minimum score for the trimmed version of the same sample where $[x]$ is the greatest integer $\leq x$, $0 \leq \tau < 0.5$, and $g=[\tau n]$, the Winsorized values (Z_i) for the scores (X_i) are given as follows: $Z = X_{(g+1)}$, if $X_i \leq X_{(g+1)}$; $Z = X_i$ if $X_{(g+1)} < X_i < X_{(n-g)}$; $Z = X_{(n-g)}$ if $X_i \geq X_{(n-g)}$. Wilcox (2003) suggested that 20% trimming is "a good choice for general use" (p. 251) and 20% trimming is applied here.

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 1. Probability of Coverage, $\alpha = 0.05$, of Yuen's & Welch's Techniques Reported in the Literature

Technique	Author(s) (Date)	n_1	n_2	σ_2/σ_1	γ_3	γ_4	$1 - \hat{\alpha}$
Yuen's	Wilcox (1994)	12	12	1	2.0	6.0	0.95
				1	3.9	42.2	0.95
	Luh & Guo (2000)	12	24	4	6.2	111.0	0.95
				1/4	6.2	111.0	0.92
	Wilcox (1994)	40	12	1	2.0	6.0	0.95
				1	3.9	42.2	0.95
		80	20	1	2.0	6.0	0.94
				1	3.9	42.2	0.95
		11	11	1	0.0	6.0	0.97
				1	0.9	1.2	0.96
Maritz-Jarrett	Charlin (1986)	25	11	1	2.0	6.0	0.96
				1	0.0	6.0	0.96
				1	0.9	1.2	0.95
		25	19	1	2.0	6.0	0.96
				1	0.0	6.0	0.96
				1	0.9	1.2	0.95
		25	25	1	2.0	6.0	0.96
				1	0.0	6.0	0.96
				1	0.9	1.2	0.95
				1	2.0	6.0	0.96
Maritz-Jarrett	Bonett & Price (2002)	15	15	9	2	6	.95
				4.7	6	110	.97
				8	1.4	3	.94
McKean-Schrader	Bonett & Price (2002)	15	15	9	2	6	.96
				4.7	6	110	.97
				8	1.4	3	.96

The technique is outlined as follows. (a) Where tau the sum of squared deviations from the Winsorized mean is given as SSD_{wi} , the Winsorized variance is estimated as $S_{wi}^2 = SSD_{wi} / (h_i - 1)$. The standard error of the trimmed mean is $S_{xi}^2 = S_{wi}^2 / h_i$. The trimmed sample size is $h_i = n_i - 2(\tau_i)$. The degrees of freedom are calculated as follows.

$$df_{\bar{x}_1 - \bar{x}_2} = \frac{[S_{x1}^2 + S_{x2}^2]^2}{[S_{x1}^2 / (h_1 - 1) + S_{x2}^2 / (h_2 - 1)]} \quad (1)$$

The $1 - \alpha/2$ percentile of Student's t-distribution provides the critical value.

Maritz-Jarrett Technique

The equations for the M-J technique were provided in Wilcox (1996). Where p_{bi} is the probability that the value of a beta random variable is between $(i-1)/n$ and i/n , $i = 1, \dots, n$. The beta probability function depends on parameter values a and b ; the parameter values for p_{bi} are given as $a = (n+1)/2$, $b = a$. The probability is obtained using the International Mathematical and Statistical Libraries (1998) function BETDF. The Harrell-Davis estimate of the population median is calculated as follows.

$$\hat{\theta} = \sum_{i=1}^n p_{bi} X_{(i)} \quad (2)$$

The variable $X_{(i)}$ is the i th ordered value of X_i . The Harrell-Davis estimate of the population median is a less biased estimate of the population median than the sample median (Wilcox, 1996, p. 73). The estimate of the variance of the median is given as follows.

$$S_{MJ}^2 = \sum_{i=1}^n p_{bi} X_{(i)}^2 - \hat{\theta}^2 \quad (3)$$

That is the variance of the median is difference between the probability of a beta random variable and the ordered value of X_i , less the square of the Harrell-Davis estimate of the population median.

McKean-Schrader Technique

The equations for the M-S technique were provided in Wilcox (2003, p. 134). The computations are given as follows as follows.

$$m = (n+1)/2 + z_{.995} \sqrt{n/4} \quad (4)$$

The estimate m is rounded to the nearest nonzero integer and $z_{.995}$ is the 99.5 percentile of the standard normal distribution. The estimate of the variance of the median is given as follows

$$S_{MS}^2 = [(X_{(n-m+1)} - X_{(m)}) / 2z_{.995}]^2 \quad (5)$$

The critical value for the M-J and M-S techniques is the $1 - \alpha/2$ percentile of the standard normal distribution. Both techniques provide confidence intervals using an estimate of the standard error of the median. Other methods of obtaining the confidence interval require the use of the bootstrap procedure (Wilcox, 1996, 2003). Modifications of the M-S variance have been recommended in Bonett and Price (2002) and Price and Bonett (2002). Yet, such modifications are not studied here.

The general form of the equation for the confidence interval for each statistic is given as follows.

$$D_{12} \mp \phi_{1-\alpha/2} SE_D \quad (6)$$

Where D_{12} represents the difference between the trimmed means or between the medians;

$\phi_{1-\alpha/2}$ represents either the two-tailed critical z or t value of a test statistic and significance level $(\alpha/2)$, SE_D is the standard error for D_{12} .

The specifications for the equation for the confidence interval were outlined in the table below. In summary, the respective measures of central tendency have a higher FSBP than the mean. Yuen's method adjusts for skewness by trimming extreme scores. The method adjusts for heteroscedasticity by the manner in which the degrees of freedom are calculated. The Maritz-Jarret technique is based on an estimate of the median that is less biased than the sample median. Further, results by Bonett and Price (2002) show that the M-S technique maintained

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 2. Equations for the Difference Estimate, Critical Value and Hypothesis for the Selected Techniques

Technique	Hypothesis	D_{12}	$\phi_{1-\alpha/2}$	SE_D
Yuen's	$H_o : \mu_{t1} = \mu_{t2}$	$\bar{X}_{t1} - \bar{X}_{t2}$	$t_{1-\alpha/2}$	$SE_{\bar{x}_{t1} - \bar{x}_{t2}} = \sqrt{S_{x1}^2 + S_{x2}^2}$
Maritz-Jarrett	$H_o : \theta_1 = \theta_2$	$X_{M1} - X_{M2}$	$z_{1-\alpha/2}$	$SEMJ_{X_{M1} - X_{M2}} = \sqrt{S_{MJ1}^2 + S_{MJ2}^2}$
McKean-Schrader	$H_o : \theta_1 = \theta_2$	$X_{M1} - X_{M2}$	$z_{1-\alpha/2}$	$SEMS_{X_{M1} - X_{M2}} = \sqrt{S_{MS1}^2 + S_{MS2}^2}$

Table 3. Descriptive Information Pertaining to Eight Real World Distributions

Distribution	μ	σ	γ_3	γ_4
Mass at Zero (MZ)	12.92	4.42	-0.03	0.31
Extreme Asymmetry-Psychometric (EAP)	13.67	5.75	1.64	1.52
Extreme Asymmetry-Achievement (EAA)	24.5	5.79	-1.33	1.11
Extreme Bimodality (EB)	2.97	1.69	-0.08	-1.70
Multimodal & Lumpy (ML)	21.15	11.9	0.19	-1.20
Digit Preference (DP)	536.95	37.64	-0.07	-0.24
Smooth Symmetric (SS)	13.19	4.91	0.01	-0.34

Note. Adapted from "A More Realistic Look at the Robustness and Type II Error Properties of the t Test to Departures From Population Normality", by S. S. Sawilowsky and R. C. Blair, 1992, *Psychological Bulletin*, 2, p. 353. Copyright 1992 by the American Psychological Association

good levels of probability of coverage under heteroscedasticity. As such, the methods were selected for consideration in the present study.

Random samples were obtained independently and with replacement using the IMSL routines: RNUND and RNSET subroutines. Each sample was obtained using the real world sampling distributions provided in Sawilowsky and Fahoome (2003). The probabilities of coverage for each interval were evaluated using samples from seven non-normal distributions that are prevalent in educational and psychological research (Micceri, 1986). The table below provides the means (μ), standard deviations (σ) and third and fourth moment estimates of skewness and kurtoses of the seven

of the eight distributions. For the purpose of study, the distribution serves as a proxy for the population. Estimates of interval-length and probability-coverage were obtained by sampling from the seven distributions. The kurtosis was adjusted so that the value for a normal distribution would be 0.00.

Sample size and standard deviation ratios of 1:1, 1:3, and 1:9 were specified. The group sizes were $n_i = 5, 15$, and 45. Sample size pairs included the following: (5,5), (5,15), (5,45), (15,15), (15,45), and (45,45). Sample size pairs were crossed with each level of heterogeneity - proportionately and disproportionately. Coverage-probabilities and interval-length were examined at the 0.05 alpha

level. As in other Monte Carlo studies, results at the 0.01 alpha level will require a larger sample size to achieve appropriate levels of Type I error.

Let k be the ratio of the standard deviation for the second group compared with the first; and X is the untransformed score; the variance for the second group was modified as follows.

$$X' = kX \quad (7)$$

Multiplying a variable by k resulted in multiplying the measure of central tendency by a factor of k . However, the Monte Carlo study requires that the variance of the second group increase while the measures of central tendency remain equal. For k greater than one, the measure of central tendency was readjusted by subtracting out the population value of central tendency (η) as follows.

$$X'' = X' - (k - 1)\eta \quad (8)$$

That is, multiplying the scores in one group by a factor increased the measure of central tendency by the same factor. Further, the measures of central tendency were adjusted to their original value. One million repetitions were done. Note that the levels of skewness, size, variance, and effect under study represent a subset of conditions that may occur in an applied situation.

The method involved (1) generating n_i random samples per group, (2) modeling heterogeneity. (3) For Yuen's technique, one had to trim and Winsorize the sample values. (4) Evaluating Equation 6 to obtain the limits of the two-sided confidence interval. (5) The location relative efficiency (LRE) for the interval width (Sawilowsky, 2002) was obtained. Welch's technique was used to calculate the LRE . The confidence interval for Welch's technique uses a separate variance estimate of the standard error. Where s_i^2 is the variance for group i ; and $s_{xi}^2 = s_i^2 / n_i$, the standard error is estimated as follows.

$$SE_{x1-x2} = \sqrt{s_{x1}^2 + s_{x2}^2} \quad (9)$$

The degrees of freedom are calculated as follows.

$$df_{wlch} = \frac{[s_{x1}^2 + s_{x2}^2]^2}{[(s_{x1}^2)^2 / (n_1 - 1) + (s_{x2}^2)^2 / (n_2 - 1)]} \quad (10)$$

The interval length for Welch's technique was used to make comparisons with that of other intervals. The equation for the LRE is obtained as follows.

$$LRE = (UL_W - LL_W) / (UL_{\{T\}} - LL_{\{T\}}) \quad (11)$$

The subscript W denotes Welch's technique and $\{T\}$ denotes either the use of Yuen's, M-J's, or M-S's technique. An LRE above one shows that the interval for the selected technique is narrower than the interval for Welch's technique; an LRE below one shows the reverse is true. Here, the standard that the comparison interval was 50% wider (narrower) than Welch's interval was adopted. (6) Estimates of probability of coverage and interval length were obtained over the number of repetitions. The average LRE served as the outcome measure for interval length comparisons. In summary, a random sample was obtained and the confidence interval was computed. The location relative efficiency was obtained and the coverage was totaled. After 1,000,000 repetitions, the average LRE and probability of coverage were obtained.

Results

Probability of Coverage

Probabilities of coverage for the M-S technique were less than 0.925 under more of the conditions than for the other techniques examined. This finding was observed for each distribution studied. Low probability of coverage occurred consistently for the inverse pairing of size and variance and for total samples sizes less than sixty, $N < 60$. However, low probability of coverage did not occur if the ratio of variances (σ_2 / σ_1) equal to

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 4. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from a Smooth Symmetric Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.956	0.961	0.955	0.885b
1	3	20	0.945	0.934	0.950	0.919b
1	9	50	0.943	0.919b	0.940	0.882b
1	1	30	0.950	0.950	0.956	0.970
1	3	60	0.949	0.946	0.951	0.963
1	1	90	0.950	0.950	0.954	0.953
3	1	10	0.948	0.945	0.941	0.867b
3	3	20	0.951	0.951	0.953	0.954
3	1/3	20	0.946	0.928	0.934	0.869b
3	9	50	0.948	0.941	0.953	0.932
3	1/9	50	0.949	0.935	0.930	0.859b
3	1	30	0.949	0.945	0.946	0.963
3	3	60	0.951	0.950	0.951	0.959
3	1/3	60	0.949	0.944	0.942	0.962
3	1	90	0.950	0.949	0.945	0.950
9	1	10	0.949	0.941	0.930	0.854b
9	3	20	0.949	0.945	0.941	0.957
9	1/3	20	0.950	0.945	0.929	0.853b
9	9	50	0.950	0.950	0.945	0.942
9	1/9	50	0.951	0.949	0.929	0.854b
9	1	30	0.950	0.945	0.939	0.958
9	3	60	0.950	0.948	0.942	0.944
9	1/3	60	0.949	0.945	0.938	0.959
9	1	90	0.950	0.948	0.941	0.945

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 5. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from an Extreme Asymmetry-Achievement Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.975	0.968	0.960	0.885b
1	3	20	0.927	0.943	0.957	0.921b
1	9	50	0.886b	0.897b	0.944	0.883b
1	1	30	0.954	0.963	0.965	0.973
1	3	60	0.939	0.945	0.959	0.966
1	1	90	0.951	0.955	0.960	0.962
3	1	10	0.899b	0.938	0.947	0.865b
3	3	20	0.951	0.956	0.961	0.954
3	1/3	20	0.881b	0.897b	0.936	0.865b
3	9	50	0.952	0.955	0.960	0.936
3	1/9	50	0.882b	0.904b	0.929	0.849b
3	1	30	0.928	0.930	0.954	0.964
3	3	60	0.949	0.952	0.958	0.964
3	1/3	60	0.923b	0.922b	0.948	0.959
3	1	90	0.944	0.942	0.952	0.954
9	1	10	0.882b	0.913b	0.930	0.842b
9	3	20	0.923b	0.925	0.949	0.955
9	1/3	20	0.883b	0.918b	0.927	0.841b
9	9	50	0.945	0.946	0.954	0.947
9	1/9	50	0.884b	0.927	0.927	0.839b
9	1	30	0.922b	0.920b	0.946	0.956
9	3	60	0.941	0.938	0.949	0.950
9	1/3	60	0.921b	0.921b	0.945	0.955
9	1	90	0.941	0.938	0.946	0.950

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 6. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from a Extreme Bimodality Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.980a	0.971	0.917b	0.648b
1	3	20	0.955	0.931	0.912b	0.698b
1	9	50	0.952	0.879b	0.913b	0.636b
1	1	30	0.950	0.946	0.901b	0.757b
1	3	60	0.948	0.933	0.905b	0.709b
1	1	90	0.949	0.948	0.916b	0.768b
3	1	10	0.954	0.942	0.868b	0.604b
3	3	20	0.958	0.952	0.866b	0.723b
3	1/3	20	0.952	0.903b	0.858b	0.602b
3	9	50	0.961	0.951	0.902b	0.770b
3	1/9	50	0.952	0.958	0.840b	0.598b
3	1	30	0.948	0.928	0.859b	0.714b
3	3	60	0.950	0.947	0.894b	0.781b
3	1/3	60	0.948	0.927	0.850b	0.698b
3	1	90	0.949	0.945	0.886b	0.778b
9	1	10	0.952	0.969	0.817b	0.595b
9	3	20	0.949	0.929	0.825b	0.671b
9	1/3	20	0.952	1.000a	0.813b	0.596b
9	9	50	0.953	0.948	0.856b	0.771b
9	1/9	50	0.952	1.000a	0.808b	0.597b
9	1	30	0.948	0.928	0.823b	0.671b
9	3	60	0.949	0.944	0.852b	0.770b
9	1/3	60	0.948	0.928	0.819b	0.672b
9	1	90	0.949	0.943	0.846b	0.766b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 7. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from a Multimodal & Lumpy Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.955	0.947	0.919b	0.779b
1	3	20	0.931	0.903b	0.919b	0.833b
1	9	50	0.921b	0.846b	0.910b	0.815b
1	1	30	0.949	0.948	0.929	0.905b
1	3	60	0.948	0.933	0.930	0.912b
1	1	90	0.950	0.949	0.946	0.934
3	1	10	0.925	0.897b	0.898b	0.769b
3	3	20	0.950	0.945	0.918b	0.888b
3	1/3	20	0.922b	0.846b	0.890b	0.770b
3	9	50	0.943	0.932	0.933	0.898b
3	1/9	50	0.923b	0.856b	0.883b	0.758b
3	1	30	0.947	0.923b	0.914b	0.896b
3	3	60	0.950	0.948	0.935	0.924b
3	1/3	60	0.946	0.918b	0.907b	0.890b
3	1	90	0.949	0.945	0.933	0.925
9	1	10	0.923b	0.865b	0.879b	0.749b
9	3	20	0.947	0.920b	0.903b	0.885b
9	1/3	20	0.924b	0.869b	0.878b	0.749b
9	9	50	0.950	0.946	0.927	0.917b
9	1/9	50	0.924b	0.875b	0.877b	0.748b
9	1	30	0.947	0.918b	0.902b	0.884b
9	3	60	0.949	0.945	0.926	0.918b
9	1/3	60	0.947	0.918b	0.900b	0.882b
9	1	90	0.949	0.944	0.925	0.917b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 8. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from a Mass at Zero Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.958	0.962	0.958	0.893b
1	3	20	0.948	0.936	0.953	0.923b
1	9	50	0.946	0.923b	0.944	0.888b
1	1	30	0.951	0.950	0.957	0.970
1	3	60	0.950	0.946	0.953	0.963
1	1	90	0.950	0.950	0.956	0.955
3	1	10	0.951	0.948	0.946	0.875b
3	3	20	0.953	0.952	0.953	0.954
3	1/3	20	0.949	0.932	0.937	0.874b
3	9	50	0.949	0.943	0.954	0.929
3	1/9	50	0.951	0.941	0.934	0.864b
3	1	30	0.950	0.945	0.945	0.962
3	3	60	0.950	0.949	0.951	0.955
3	1/3	60	0.950	0.943	0.940	0.959
3	1	90	0.950	0.947	0.947	0.939
9	1	10	0.952	0.945	0.934	0.859b
9	3	20	0.951	0.945	0.939	0.954
9	1/3	20	0.953	0.950	0.932	0.859b
9	9	50	0.951	0.949	0.945	0.937
9	1/9	50	0.954	0.953	0.932	0.860b
9	1	30	0.950	0.943	0.937	0.955
9	3	60	0.950	0.947	0.944	0.942
9	1/3	60	0.951	0.943	0.936	0.956
9	1	90	0.950	0.947	0.945	0.945

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 9. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from an Extreme Asymmetry-Psychometric Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.990a	0.985a	0.974	0.894b
1	3	20	0.955	0.984a	0.969	0.920b
1	9	50	0.893b	0.971	0.960	0.888b
1	1	30	0.962	0.988a	0.987a	0.983a
1	3	60	0.937	0.964	0.985a	0.980a
1	1	90	0.951	0.968	0.980a	0.973
3	1	10	0.907b	0.984a	0.969	0.888b
3	3	20	0.952	0.962	0.963	0.941
3	1/3	20	0.878b	0.973	0.967	0.896b
3	9	50	0.968	0.972	0.932	0.864b
3	1/9	50	0.878b	0.972	0.964	0.872b
3	1	30	0.920b	0.913b	0.972	0.974
3	3	60	0.949	0.952	0.937	0.935
3	1/3	60	0.913b	0.877b	0.953	0.953
3	1	90	0.940	0.918b	0.882b	0.845b
9	1	10	0.876b	0.984a	0.963	0.866b
9	3	20	0.912b	0.900b	0.937	0.930
9	1/3	20	0.878b	0.985a	0.962	0.863b
9	9	50	0.940	0.933	0.848b	0.764b
9	1/9	50	0.881b	0.990a	0.960	0.858b
9	1	30	0.911b	0.867b	0.927	0.934
9	3	60	0.937	0.905b	0.813b	0.762b
9	1/3	60	0.911b	0.866b	0.921b	0.929
9	1	90	0.936	0.902b	0.798b	0.739b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table10. Probabilities of Coverage for the Welch, Yuen, Maritz-Jarrett, & McKean-Schrader Techniques for Equal & Unequal Sample Sizes & Variances when Sampling is from a Digit Preference Distribution

σ_2/σ_1	n_2/n_1	N	Welch	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.955	0.960	0.955	0.888b
1	3	20	0.945	0.935	0.951	0.922b
1	9	50	0.943	0.920b	0.941	0.887b
1	1	30	0.951	0.950	0.958	0.973
1	3	60	0.949	0.946	0.955	0.967
1	1	90	0.950	0.949	0.958	0.966
3	1	10	0.948	0.945	0.941	0.869b
3	3	20	0.952	0.951	0.954	0.956
3	1/3	20	0.947	0.928	0.934	0.871b
3	9	50	0.948	0.941	0.952	0.935
3	1/9	50	0.949	0.934	0.931	0.859b
3	1	30	0.949	0.945	0.948	0.965
3	3	60	0.950	0.949	0.947	0.958
3	1/3	60	0.949	0.945	0.943	0.962
3	1	90	0.950	0.948	0.938	0.946
9	1	10	0.949	0.938	0.930	0.855b
9	3	20	0.949	0.946	0.942	0.959
9	1/3	20	0.950	0.941	0.929	0.855b
9	9	50	0.950	0.949	0.932	0.934
9	1/9	50	0.951	0.944	0.928	0.853b
9	1	30	0.950	0.945	0.940	0.961
9	3	60	0.950	0.948	0.927	0.936
9	1/3	60	0.950	0.945	0.939	0.960
9	1	90	0.950	0.947	0.925	0.933

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 11. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from a Smooth Symmetric Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.803	0.882	1.140b
1	3	20	0.904	0.875	1.002b
1	9	50	0.946b	0.941	1.157b
1	1	30	0.906	0.782	0.712
1	3	60	0.916	0.794	0.732
1	1	90	0.913	0.760	0.743
3	1	10	0.856	0.955	1.230b
3	3	20	0.890	0.812	0.794
3	1/3	20	0.922	0.994	1.245b
3	9	50	0.902	0.804	0.888
3	1/9	50	0.894	1.020	1.301b
3	1	30	0.917	0.823	0.741
3	3	60	0.912	0.770	0.738
3	1/3	60	0.919	0.843	0.756
3	1	90	0.916	0.783	0.763
9	1	10	0.868	1.021	1.312b
9	3	20	0.916	0.843	0.763
9	1/3	20	0.853	1.031	1.320b
9	9	50	0.910	0.784	0.786
9	1/9	50	0.834	1.035	1.329b
9	1	30	0.918	0.851	0.759
9	3	60	0.916	0.792	0.768
9	1/3	60	0.918	0.855	0.761
9	1	90	0.916	0.796	0.774

a. $1 - \hat{\alpha} > 0.975$ b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 12. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from an Extreme Asymmetry-Achievement Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.895	0.898	1.207b
1	3	20	0.999	0.893	1.046b
1	9	50	1.069b	0.941	1.188b
1	1	30	0.988	0.854	0.773
1	3	60	1.008	0.879	0.806
1	1	90	0.995	0.867	0.839
3	1	10	0.954	0.948	1.271b
3	3	20	0.974	0.873	0.870
3	1/3	20	1.053b	0.979	1.265b
3	9	50	0.982	0.863	0.972
3	1/9	50	1.038b	1.007	1.331b
3	1	30	1.020	0.904	0.810
3	3	60	0.995	0.869	0.825
3	1/3	60	1.038b	0.936	0.836
3	1	90	1.007	0.899	0.869
9	1	10	0.986b	1.006	1.346b
9	3	20	1.026	0.929	0.845
9	1/3	20	0.981b	1.019	1.355b
9	9	50	0.993	0.888	0.894
9	1/9	50	0.952	1.025	1.370b
9	1	30	1.040b	0.946	0.840
9	3	60	1.011	0.910	0.877
9	1/3	60	1.041b	0.953	0.845
9	1	90	1.013	0.920	0.888

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 3. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from a Extreme Bimodality Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.564	0.746b	1.281b
1	3	20	0.631	0.673b	1.043b
1	9	50	0.650b	0.725b	1.124b
1	1	30	0.623	0.463b	0.626b
1	3	60	0.622	0.428b	0.563b
1	1	90	0.608	0.341b	0.430b
3	1	10	0.594	0.830b	1.408b
3	3	20	0.617	0.527b	0.717b
3	1/3	20	0.608b	0.850b	1.405b
3	9	50	0.623	0.471b	0.626b
3	1/9	50	0.575	0.872b	1.442b
3	1	30	0.630	0.495b	0.648b
3	3	60	0.613	0.377b	0.471b
3	1/3	60	0.633	0.498b	0.644b
3	1	90	0.608	0.363b	0.442b
9	1	10	0.568	0.891b	1.499b
9	3	20	0.631	0.519b	0.668b
9	1/3	20	0.554a	0.895b	1.503b
9	9	50	0.610	0.392b	0.471b
9	1/9	50	0.547a	0.897b	1.509b
9	1	30	0.629	0.517b	0.662b
9	3	60	0.608	0.382b	0.453b
9	1/3	60	0.629	0.519b	0.662b
9	1	90	0.607	0.383b	0.450b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 14. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from a Multimodal & Lumpy Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.715	0.839b	1.173b
1	3	20	0.802b	0.790b	0.995b
1	9	50	0.884b	0.862b	1.115b
1	1	30	0.753	0.617	0.658b
1	3	60	0.758	0.607	0.634b
1	1	90	0.747	0.538	0.550
3	1	10	0.800b	0.920b	1.273b
3	3	20	0.750	0.676b	0.740b
3	1/3	20	0.911b	0.949b	1.278b
3	9	50	0.768	0.651	0.734b
3	1/9	50	0.938b	0.978b	1.324b
3	1	30	0.763b	0.658b	0.682b
3	3	60	0.748	0.566	0.579b
3	1/3	60	0.767b	0.674b	0.686b
3	1	90	0.746	0.564	0.565
9	1	10	0.880b	0.988b	1.352b
9	3	20	0.765b	0.686b	0.701b
9	1/3	20	0.913b	0.996b	1.358b
9	9	50	0.748	0.586	0.593b
9	1/9	50	0.893b	1.001b	1.365b
9	1	30	0.767b	0.688b	0.696b
9	3	60	0.746	0.580	0.574b
9	1/3	60	0.767b	0.691b	0.697b
9	1	90	0.745	0.582	0.573b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 15. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from a Mass at Zero Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.834	0.892	1.143b
1	3	20	0.937	0.889	1.015b
1	9	50	0.974b	0.949	1.162b
1	1	30	0.948	0.818	0.743
1	3	60	0.958	0.829	0.765
1	1	90	0.956	0.797	0.781
3	1	10	0.882	0.960	1.228b
3	3	20	0.929	0.840	0.822
3	1/3	20	0.939	0.998	1.246b
3	9	50	0.940	0.828	0.914
3	1/9	50	0.904	1.024	1.301b
3	1	30	0.959	0.859	0.772
3	3	60	0.955	0.807	0.775
3	1/3	60	0.962	0.878	0.788
3	1	90	0.958	0.821	0.803
9	1	10	0.882	1.024	1.311b
9	3	20	0.957	0.878	0.795
9	1/3	20	0.864	1.035	1.320b
9	9	50	0.952	0.819	0.825
9	1/9	50	0.847	1.038	1.329b
9	1	30	0.961	0.887	0.791
9	3	60	0.959	0.830	0.809
9	1/3	60	0.960	0.891	0.793
9	1	90	0.959	0.835	0.817

a. $1 - \hat{\alpha} > 0.975$ b. $1 - \hat{\alpha} < 0.925$

ASSESSING THE MEDIAN AND TRIMMED MEAN

Table 16. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from an Extreme Asymmetry-Psychometric Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	1.008a	0.894	1.277b
1	3	20	1.156a	0.935	1.150b
1	9	50	1.265	0.977	1.328b
1	1	30	1.225a	1.047a	0.898a
1	3	60	1.253	1.151a	1.026a
1	1	90	1.175	1.255a	1.283
3	1	10	1.080a	0.938	1.336b
3	3	20	1.171	0.997	1.027
3	1/3	20	1.238	0.975	1.329b
3	9	50	1.117	1.008	1.256b
3	1/9	50	1.243	0.997	1.405b
3	1	30	1.350b	1.147	0.979
3	3	60	1.191	1.197	1.148
3	1/3	60	1.456b	1.254	1.079
3	1	90	1.239b	1.360b	1.404b
9	1	10	1.164a	0.987	1.406b
9	3	20	1.377b	1.178	1.073
9	1/3	20	1.171a	1.001	1.415b
9	9	50	1.177	1.216b	1.378b
9	1/9	50	1.131a	1.008	1.439b
9	1	30	1.508b	1.270	1.083
9	3	60	1.268b	1.391b	1.412b
9	1/3	60	1.544b	1.307b	1.120
9	1	90	1.293b	1.488b	1.548b

a. $1 - \hat{\alpha} > 0.975$

b. $1 - \hat{\alpha} < 0.925$

Table 17. Length Ratios for the Yuen, Maritz-Jarrett, & McKean-Schrader Techniques Compared with Welch's Technique for Equal & Unequal Sample Sizes & Variances when Sampling is from a Digit Preference Distribution

σ_2/σ_1	n_2/n_1	N	Yuen	Maritz-Jarrett	McKean-Schrader
1	1	10	0.812	0.886	1.139b
1	3	20	0.915	0.882	1.002b
1	9	50	0.964b	0.950	1.157b
1	1	30	0.915	0.797	0.716
1	3	60	0.925	0.815	0.739
1	1	90	0.919	0.798	0.756
3	1	10	0.872	0.958	1.227b
3	3	20	0.900	0.824	0.797
3	1/3	20	0.953	0.998	1.243b
3	9	50	0.912	0.827	0.896
3	1/9	50	0.945	1.025	1.298b
3	1	30	0.927	0.839	0.745
3	3	60	0.919	0.803	0.749
3	1/3	60	0.930	0.861	0.760
3	1	90	0.922	0.825	0.776
9	1	10	0.905	1.024	1.309b
9	3	20	0.926	0.859	0.767
9	1/3	20	0.905	1.035	1.317b
9	9	50	0.917	0.822	0.798
9	1/9	50	0.881	1.039	1.326b
9	1	30	0.929	0.869	0.763
9	3	60	0.923	0.834	0.781
9	1/3	60	0.929	0.873	0.766
9	1	90	0.923	0.841	0.788

a. $1 - \hat{\alpha} > 0.975$ b. $1 - \hat{\alpha} < 0.925$

nine with the ratio of sample sizes (n_2/n_1) equal to three or one, or for total sample sizes greater than sixty ($N \geq 60$). Results were presented in the tables. Where absolute kurtosis was above 1.15, the probability of coverage for each technique was less than 0.925. Low probability of coverage occurred consistently for the inverse pairing of size and variance and for total sample sizes less than sixty ($N < 60$).

Location Relative Efficiency

The conditions of the inverse pairing of size and variance and total sample sizes less than 60 that resulted in probability of coverage less than 0.925 also results in *LREs* at or above 1.25 using the M-S technique. That is, the interval length for the M-S technique was 80% of the length for Welch's technique. The results were presented in the tables below. If absolute kurtosis was less than 1.15, the interval length for Yuen's technique and the interval length for the M-J technique approximated the length for Welch's technique. The length ratios were within the following range: 0.75-1.10. Further, the M-J technique displayed appropriate levels of coverage for each size and variance pairing.

If absolute kurtosis exceeded 1.15, specified results were observed for the EB distribution and for the multimodal lumpy distribution. (a) Under an EB distribution, the following was observed. The length for Yuen's technique exceeded that for Welch's technique by 50% or more. Length ratios for Yuen's technique were below 0.66. Yuen's technique also displayed appropriate levels of coverage for more of the size and variance conditions than either the Maritz-Jarrett or McKean-Schrader techniques. (b) Under a multimodal lumpy distribution, the following was observed. If total sample sizes exceeded 50, were equal or were proportional to variances, the width for the M-J technique was 50% wider than for Welch's technique. *LREs* for the M-J technique were less than 0.66. On the other hand, the interval lengths for Yuen's technique approximated that for Welch's technique. *LREs* were within the range 0.715-0.940. For both Yuen's and the M-J techniques, the probability of coverage was still less than 0.925 under several size and variance conditions. Low probability of coverage

occurred consistently with the inverse pairing of size and variance and for total sample sizes less than 60.

Conclusion

The purpose of the present study was to assess the probability of coverage and interval length of selected statistical techniques that have a higher FSBP than the mean and appropriate levels of probability of coverage when using Bradley's (1978) criterion. The techniques were examined using real education and psychology datasets (Sawilowsky & Fahoome, 2003, Sawilowsky & Blair, 1992). Welch's test appears to be robust to minor violations involving heteroscedasticity. Welch's test exhibited appropriate coverage for the smooth symmetric, mass at zero, digit preference, and extreme bimodal distributions. In general, the M-S technique exhibits narrow interval lengths that do not provide accurate coverage. The M-J technique does not perform well when kurtosis is at or below -1.25. However, it does perform well otherwise. Yuen's technique does perform well when kurtosis is below -1.25. However, its performance approximated that of the M-J technique under an ML distribution. A tradeoff was observed between coverage and length for Yuen's technique. Adequate coverage is often observed with wider interval lengths and vice versa.

References

References marked with an asterisk were included in Table 1.

Bonett, D. G., & Price, R. M. (2002). Statistical inference for a linear function of medians: Confidence intervals, hypothesis testing, and sample size requirements. *Psychological Methods*, 7, 370-383.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology*, 31, 144-152.

Bunner, J. M. (2003). *Forming a bracketed interval around the trimmed mean: Alternatives to Sw* (Doctoral Dissertation, Wayne State University, 2003). Dissertation Abstracts International, 64, 6147.

- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Hinkle, D. E., Wiersma, W., & Jurs, S. G. (1998). *Applied statistics for the behavioral sciences*. Boston, MA: Houghton Mifflin Company.
- Hoening, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *American Statistician*, 55, 19-23.
- International Mathematical and Statistical Libraries (1998). *IMSL F90 MP library 3.0*. Incline Village, NV: Lahey Computer Systems, Inc.
- Knapp, T. R. (1998). Comments on statistical significance testing articles. *Research in the Schools*, 5, 39-41.
- Knapp, T. R., & Sawilowsky, S. S. (2001). Constructive criticism of methodological and editorial practices. *Journal of Experimental Education*, 70, 65-79.
- Lahey Computer Systems. (1995-2000). Essential lahey FORTRAN 90. Incline Village, NV: Lahey Computer Systems, Inc.
- *Luh, W., & Guo, J. (2000). Johnson's transformation two-sample trimmed t and its bootstrap method for heterogeneity and non-normality. *Journal of Applied Statistics*, 27, 965-973.
- Mends-cole, S. J. (2006). Coverage & interval length for Welch's and Yuen's procedures for shift in location and change in scale for (Un)equal sample sizes (Doctoral Dissertation, Wayne State University, 2006). Dissertation Abstracts International.
- Price, R. M., & Bonett, D. G. (2002). Distribution-free confidence intervals for difference and ratio of medians. *Journal of Statistical Computation and Simulation*, 72, 119-124.
- Sawilowsky, S. S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures from population normality. *Psychological Bulletin*, 111, 352-360.
- Sawilowsky, S. S., & Fahoome, G. (2003). *Statistics through Monte Carlo experimentation with FORTRAN*. Oak Park, MI: JMASM, Inc.
- Sawilowsky, S. S. (2002). A measure of relative efficiency for location of a single sample. *Journal of Modern & Applied Statistical Methods*, 1, 52-60.
- *Wilcox, R. R. (1994). Some results on the Tukey-McLaughlin and Yuen methods for trimmed means when distributions are skewed. *Biometrical Journal*, 3, 259-273.
- Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.
- Wilcox, R. R. (2001). *Fundamentals of modern statistical methods*. New York, NY: Springer-Verlag.
- Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. San Diego, CA: Academic Press.
- *Wilcox, R. R., & Charlin, V. L. (1986). Comparing medians: A Monte Carlo study. *Journal of Educational Statistics*, 11, 263-274.
- Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594-604.
- Zumbo, B. D., & Hubley, A. M. (1998). A note on misconceptions concerning prospective and retrospective power. *The Statistician*, 47, 385-388.

Tests for Independence in Two-Way Contingency Tables with Small Samples

Stephen Sharp
University of Edinburgh

When testing the null hypothesis of independence in a two-way contingency table, the likelihood ratio test statistic is approximately distributed as Chi-squared d for large sample sizes (N) but may not be for small samples. This paper presents expressions which match the mean of the statistic to Chi-squared d as far as N^{-1} and N^{-2} , derives a method of estimating the expressions from observed data and evaluates them using Monte Carlo simulations. It is concluded that using appropriate dividing factors, rejection rates after matching are more accurate than for either the unadjusted likelihood ratio statistic or the Pearson approximation which is the main alternative statistic. Minimum cell frequencies necessary for high test accuracy are smaller than those commonly given in textbooks.

Key words: Contingency tables, likelihood ratios, small samples, dividing factors.

Introduction

A common requirement in social science research is to test the null hypothesis of independence between the two axes of a contingency table. It is well known that this can be tested either by using Pearson's Chi-squared approximation based on squares of differences between observed and expected values, or by the likelihood ratio test statistic (LR) originally proposed by Neyman and Pearson (1928). Many widely used statistical packages like SPSS give both statistics. Both tend asymptotically to a Chi-squared d distribution as sample size increases. Tabachnick and Fidell (2004, p. 251) pointed out that from a theoretical point of view, LR is preferable because it is available for testing overall fit, screening, and testing for differences among hierarchical models. However LR has the relative disadvantage that it converges to Chi-squared d more slowly than

the Pearson test and that for small samples it has values which are stochastically larger than its asymptote (i.e. it errs in the 'wrong' direction). This is because LR is based on the function $\sum p \log p$ where the summation goes over a complete set of probabilities, and as this is a downward concave function, replacing probabilities by their estimates leads to bias in estimating the sum. A similar effect has been noted where the same function is used to estimate entropy in physical systems (Victor 2000).

Starting in the 1950s, statisticians have tried to find ways of adjusting the LR test to speed up its rate of convergence to Chi-squared and hence overcome its main limitation. The classic papers of Bartlett (1954) and Lawley (1957) developed a general method which applies to all continuous likelihood functions. The Bartlett-Lawley adjustment takes the form of a number, which, if used as a divisor for the LR statistic, matches all its moments to those of Chi-squared as far as terms in N^{-1} , where N is the sample size, thus accelerating the convergence. However Frydenberg and Jensen (1989) cast doubt on whether Lawley's method is effective at all when applied to discrete data (of which contingency tables are an example). They point out that the Lawley method assumes that the LR statistic can be written as a function of a continuous random variable, which is not the case with contingency tables with discrete

Stephen Sharp is a Visiting Scholar at the Centre for Educational Leadership in the Moray House School of Education. His research interests include educational assessment, evaluation and monitoring; teacher careers and professional development; and statistical applications in educational research. E-mail him at stephen.sharp@ed.ac.uk.

cell frequencies. The method however may still be valid to the extent that discrete data approximates continuous data, which will increasingly be the case as the number of cells in the table increases. Frydenberg and Jensen presented evidence that the method is seriously in error for one-dimensional frequency tables with three and four categories, where the data is at its most discrete, as it were. But other writers have argued that this view is overly pessimistic. Pierce and Peters (1992) explored one-parameter functions of exponential families, finding that excellent approximations can be obtained from simple adjustments to the signed square root of the likelihood ratio statistic with one degree of freedom. They did not however consider alternative hypotheses of a more generalized nature, as the present paper does.

Williams (1976, 1978) derived first-order adjustments for the LR statistic for one-way, two-way and five-way contingency tables, though he did this not by using the Bartlett-Lawley method directly but by expanding $n \log n$ (where N is an observed cell frequency) as a Taylor series around its mean. Also, he did not offer any empirical evaluation of the expressions he derived. Subsequently Smith *et al* (1981) also used a Taylor expansion around the mean and, by taking more terms in the series, derived a second-order expression which matches the first moment of LR to its asymptotic value as far as terms in N^{-2} and all others as far as terms in N^{-1} . Smith *et al* considered only one-way tables with the null hypothesis of equal probabilities but produced evidence that their more accurate adjustment did indeed model upper cut-offs more accurately than that of Williams. Bayo Lawal (1984) showed that the Pearson test performs as well as the Williams-adjusted LR for one-way tables with 3 and 4 cells, though he did not consider the adjustment of Smith *et al*.

The aim of the present paper is to extend Williams' expression for two-way tables from first order (as far as N^{-1}) to second order (as far as N^{-2}) levels of accuracy for the mean of the test statistic. Put another way, it is to extend Smith *et al*'s second-order expression from one-way to two-way tables. The paper also presents Monte Carlo simulations to evaluate various

adjustments to the LR statistic against each other and against the Pearson test and considers various practical issues concerning the implementation of the adjustments.

Methodology

First-order adjustments for two-way tables

To evaluate the Taylor series which results from expanding $n \log n$, it is necessary to assume how N is distributed. Williams took the Poisson distribution while Smith *et al* used the multinomial. The latter is used throughout this paper for consistency. The distributions are closely linked and lead to the same answer for the first-order divisor, which Williams showed to be given by the expression

$$1 + \frac{\left(\sum_i r_i^{-1} - 1 \right) \left(\sum_j c_j^{-1} - 1 \right)}{6N(r-1)(c-1)} \quad (1)$$

where the table has r rows and c columns with marginal probabilities r_i ($i = 1, 2, \dots, r$) and c_j ($j = 1, 2, \dots, c$). Williams pointed out that the effect of the adjustment will be minimized where all the r_i equal $1/r$ and all the c_j equal $1/c$. In this case, $\sum_i r_i^{-1} = r^2$ and $\sum_j c_j^{-1} = c^2$ so the above expression can be written simply as

$$1 + \frac{(r+1)(c+1)}{6N} \quad (2)$$

This is undoubtedly safe but perhaps the adjustment may be made more accurate by estimating the sums of the reciprocals of the probabilities from the data. Neither Williams nor Smith *et al* considered the practicalities of doing this, the former because he attempted no empirical validation of the expression and the latter because they considered only the null hypothesis of uniformity where the parameters are known and do not have to be estimated.

The naive estimate of r_i^{-1} (an analogous argument applies to the column probabilities) is simply the reciprocal of its maximum likelihood estimate ie R_i / N , where R_i is the i th row total.

However N/R_i is not an unbiased estimate of r_i^{-1} . In fact the expected value of N/R_i is undefined if R_i follows a Poisson or multinomial distribution as there is a finite probability that R_i equals zero. In practice when using contingency tables, rows and columns with no observations at all are deleted from the analysis, the degrees of freedom being reduced accordingly. For present purposes however the problem is to estimate r_i^{-1} where R_i might be zero. To do this, we consider the expected value of $(R_i+1)^{-1}$. From the density function of the binomial distribution, this is given by

$$E \frac{1}{R_i+1} = \sum_{R_i=0}^N \frac{N!}{(R_i+1)!(N-R_i)!} r_i^{R_i} (1-r_i)^{N-R_i}$$

We make the binomial series complete again by multiplying by $r_i(N+1)$ and adding in a term for $R_i = -1$. Rearranging yields

$$E \left(\frac{N+1}{R_i+1} \right) = \frac{1}{r_i} [1 - (1-r_i)^{N+1}] \quad (3)$$

The left hand side is an underestimate of r_i^{-1} , the error being $(1-r_i)^{N+1}$. However this is less than 5% if the expected value of R_i is around three and less than 1% if it is around five, values which should be exceeded comfortably by sample sizes used in practice in research. An analogous argument leads to $(N+1)/(C_j+1)$ as an estimate of c_j^{-1} .

Second-order adjustments for two-way tables

This is achieved in the same way as the first-order adjustments except that more terms are taken from the Taylor series. The expression derived by Smith *et al* for the second-order divisor for one-way tables was

$$1 + \frac{\sum p_i^{-1} - 1}{6N(k-1)} + \frac{\sum p_i^{-2} - \sum p_i^{-1}}{6N^2(k-1)} \quad (4)$$

where there are k categories with probabilities p_i . The same method can be applied to the null hypothesis of independence in a two-way table rather than that of specified p -values in a one-way table. The resulting algebra is laborious but straightforward. It leads to the rather ungainly expression

$$1 + \frac{\left(1 - \frac{1}{N}\right) \left(\sum_i r_i^{-1} - 1\right) \left(\sum_j c_j^{-1} - 1\right)}{6N(r-1)(c-1)} + \frac{\left(\sum_i r_i^{-2} - 1\right) \left(\sum_j c_j^{-2} - 1\right)}{6N^2(r-1)(c-1)} \quad (5)$$

which is clearly a combination of (1) and (4). In the second-order case, the use of the Poisson or multinomial assumption makes a difference. The above version is the multinomial one. The term $1-1/N$ in the numerator of the middle part of (5) disappears in the Poisson version, but in practice the difference between the two will be negligible if N has a value which is reasonable for research purposes.

Again there is a 'safe' version of this based on the assumption that all the r_i equal $1/r$ and all the c_j equal $1/c$. The result is

$$1 + \frac{(1-1/N)(r+1)(c+1)}{6N} + \frac{(r^2+r+1)(c^2+c+1)}{6N^2} \quad (6)$$

Again, the $1-1/N$ term is absent if the Poisson distribution is assumed. Following an argument analogous to that used above, we estimate r_i^{-2} by considering the expected value of the reciprocal of $(R_i+1)(R_i+2)$:

$$E \frac{1}{(R_i+1)(R_i+2)} = \sum_{R_i=0}^N \frac{N!}{(R_i+2)!(N-R_i)!} r_i^{R_i} (1-r_i)^{N-R_i}$$

This time we complete the binomial series by multiplying by $r_i^2(N+1)(N+2)$

and adding in terms for $R_i = -1$ and $R_i = -2$. Rearranging yields

$$E \frac{(N+1)(N+2)}{(R_i+1)(R_i+2)} = \frac{1}{r_i^2} \left\{ 1 - (1-r_i)^{N+1} [1 + r_i(N+1)] \right\} \quad (7)$$

The extent of the underestimation of r_i^{-2} is greater than for r_i^{-1} but it is still less than 5% if the expected value of R_i is at least five and less than 1% if it is at least seven, and these are also values which should be exceeded comfortably by sample sizes used in practice in research. An analogous argument leads to $(N+1)(N+2)/[(C_j+1)(C_j+2)]$ as an estimate of c_j^{-2} . However it is not certain that a second-order expression based on estimated parameters will be successful in improving the accuracy of the method. Victor (2000) found that this approach did not always lead to greater accuracy when trying to derive improved estimates for entropy in physical systems. The figures reported in the next section throw light on the accuracy of the various adjustments.

Results

Monte Carlo methods were used to assess the accuracy of six different tests which are summarized in Table 1. The choice of which sort of simulated data to use is inevitably to some extent arbitrary. The choice used here is based on the advice offered by most statistical text books (e. g., Tabachnick & Fidell 2004, p. 223) that the Pearson test should not be used unless all expected cell frequencies in the table are greater than one and not more than one-fifth of them are less than five. Tabachnick and Fidell do not give the source of this advice and there seems to be no corresponding advice for the *LR* test.

All the contingency tables used in the simulations had five columns with the number of rows being two, three, four and five. In each row the expected value of the first cell was one while all other cells had the expected value M where M

had the values two, three, four and five. Thus in all cases, one-fifth of cells have an expected value of one and all other cells have an expected value of M . The case where M equals five is the criterion case for the advice given in statistics texts. The aim is to investigate whether any of the adjusted tests perform well with values of M less than five.

For each of the 16 (four numbers of rows by four values of M) versions of the table, 10,000 sets of data were simulated where the null hypothesis was true. The results are contained in table 2. For ease of interpretation, some of the entries in this table are in bold face. If the actual and nominal rejection rates for a test are the same, the percentage of rejections at a level of significance p (e. g., 0.05) has an expected value of $100p$ and variance $p(1-p)$. Entries in the tables where the observed percentage is within two standard deviations of the nominal percentage are in bold type. This is a stringent criterion as the only deviation which it allows from the nominal rejection levels is that expected on the basis of sampling error.

The main comparison is between tests 2 to 6 (i. e., the various *LR* tests). Test 1, the Pearson approximation, acts as a benchmark. It is immediately clear from table 2 that test 2 (the unadjusted *LR* test) is seriously in error, the rejection rate being well above the nominal rate for all levels of significance, especially the less stringent ones. The smallest dividing factor is test 3, the *LR* test with first order adjustment based on equal marginal parameters. This is a marked improvement on test 2 but still has a tendency to over-reject slightly at the 10% and 5% levels. Larger adjustments are provided by tests 4 and 5 and these have higher concentrations of accurate rejection rates. There is little to choose between them except at the very smallest sample size where $M=2$ and each row has an expected frequency of just nine. Here, test 4 over rejects slightly at the 10% level but is more accurate than test 5 at the 5% and 1% levels (the levels most often used in social science research). Test 6, the second-order correction with estimated parameters, has the largest adjustment of all but appears to be a step too far, as it were. Its performance is similar to test 1 (the Pearson approximation), i. e., safe but

conservative, especially at the more stringent significance levels, but not as accurate as tests 4 or 5.

It might seem counterintuitive that test 6, which has the strongest theoretical rationale, is not the most accurate but in fact this is not so surprising. As Smith et al. (1981) pointed out, the use of a scaling factor to match moments is by its very nature a fairly crude device whose effects at a fine level of detail may not always match closely with theoretical expectations. Also, the adjustments are designed to match the first moment of the LR statistic with its asymptotic mean, and the mean values (not reported here) observed in the simulations do indeed show that test 6 usually produces the mean closest to the number of degrees of freedom. However the criterion used here (and the one in which test users are interested) is the accuracy with which each test models not the mean but the upper cut-off scores and this depends on characteristics of the distribution other than the mean.

Conclusion

On the basis of the arguments and data reported above, these conclusions are offered:

- (i) the unadjusted LR test should not be used for small samples;
- (ii) the Pearson approximation is safe but conservative;
- (iii) the view of Frydenberg and Jensen is overly pessimistic in the context of two-way contingency tables where the use of rescaling factors can result in improved test accuracy;
- (iv) if a second-order adjustment assuming equal marginal probabilities is used based on (6) above, then at the 10% and 5% levels of significance, accurate rejection rates are achieved where all expected values are at least one and not more than one-fifth are less than three; and
- (v) if a first-order adjustment is used with marginal probabilities estimated from the data using the method based on (1) and (3) above, accurate rejection rates are achieved where all expected values are at

least one and not more than one-fifth are less than two (for the 5% level of significance) or three (for the 1% level).

References

- Bartlett, M. S. (1954). A note on the multiplying factors in various Chi-squared approximations. *Proceedings of the Cambridge Philosophical Society*, 47, 86-95.
- Bayo Lawal, H. (1984). Comparisons of the X^2 , Y^2 , Freeman-Tukey and Williams' improved G^2 test statistics in small samples of one-way multinomials. *Biometrika*, 71, 415-418.
- Frydenberg, M. & Jensen, J. L. (1989). Is the 'improved likelihood ratio statistic' really improved in the discrete case? *Biometrika*, 76, 655-661.
- Lawley, D. N. (1957). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, 43, 295-303.
- Neyman, J. & Pearson, E. S. (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, 20, 175-240 and 263-294.
- Pierce, D. A. & Peters, D. (1992) Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society*, B54, 701-737.
- Smith, P. J., Rae, D. S., Manderscheid, R. W. & Silbergeld, S. (1981). Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness of fit. *Journal of the American Statistical Association*, 76, 737-740.
- Tabachnick, B. G. & Fidell, L. S. (2004). *Using multivariate statistics*. (4th edition) New York: Allyn and Bacon.
- Victor, J. D. (2000). Asymptotic bias in information estimates and the exponential (Bell) polynomials. *Neural Computation*, 12, 2797-2804.
- Williams, D. A. (1976). Improved likelihood ratio tests for complete contingency tables. *Biometrika*, 63, 33-37.
- Williams, D. A. (1978). Amendments and corrections: improved likelihood ratio tests for complete contingency tables. *Biometrika*, 65, 77.

Table 1: Adjustments to LR tests and their divisors

Test Description	Divisor
1 Pearson approximation.	-
2 Unadjusted LR test.	-
3 LR test with first order adjustment and equal marginal parameters.	expression (2)
4 LR test with first order adjustment and estimated marginal parameters.	expressions (1) and (3)
5 LR test with second order adjustment and equal marginal parameters.	expression (6)
6 LR test with second order adjustment and estimated marginal parameters.	expressions (5) and (7)

TESTS FOR INDEPENDENCE IN TWO-WAY CONTINGENCY TABLES

Table 2: Rejection rates for six tests and various tables sizes and cell frequencies. Bold entries are within sampling variation of the nominal rate.

	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 1	2	2	8.58	3.07	0.12	0.00	Test 2	2	2	18.99	10.42	1.71	0.06
Test 1	3	2	8.56	3.73	0.52	0.03	Test 2	3	2	19.78	10.51	1.87	0.13
Test 1	4	2	8.98	4.34	0.67	0.02	Test 2	4	2	20.97	11.65	2.30	0.14
Test 1	5	2	9.29	4.15	0.59	0.03	Test 2	5	2	22.88	12.57	2.56	0.20
Test 1	2	3	9.06	3.74	0.43	0.00	Test 2	2	3	16.31	8.99	2.06	0.21
Test 1	3	3	9.60	4.07	0.55	0.02	Test 2	3	3	17.81	9.93	1.88	0.12
Test 1	4	3	9.44	4.47	0.85	0.08	Test 2	4	3	19.20	10.58	2.48	0.23
Test 1	5	3	9.25	4.49	0.80	0.08	Test 2	5	3	19.00	10.27	2.48	0.29
Test 1	2	4	8.59	3.96	0.52	0.03	Test 2	2	4	14.10	7.52	1.87	0.24
Test 1	3	4	9.11	4.19	0.76	0.05	Test 2	3	4	15.42	8.05	2.04	0.24
Test 1	4	4	9.12	4.30	0.80	0.04	Test 2	4	4	15.90	8.69	1.90	0.21
Test 1	5	4	9.04	4.12	0.83	0.08	Test 2	5	4	16.66	8.73	2.01	0.27
Test 1	2	5	9.47	4.24	0.74	0.02	Test 2	2	5	13.64	7.50	1.83	0.23
Test 1	3	5	9.89	4.48	0.61	0.04	Test 2	3	5	14.57	7.76	1.66	0.18
Test 1	4	5	9.80	4.78	0.95	0.08	Test 2	4	5	14.84	8.36	1.88	0.24
Test 1	5	5	9.04	4.03	1.00	0.08	Test 2	5	5	14.81	7.70	1.98	0.24
	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 3	2	2	11.55	5.04	0.49	0.00	Test 4	2	2	11.96	5.23	0.53	0.00
Test 3	3	2	11.01	4.80	0.60	0.00	Test 4	3	2	10.98	4.78	0.62	0.00
Test 3	4	2	11.37	5.08	0.61	0.01	Test 4	4	2	11.02	4.84	0.60	0.01
Test 3	5	2	11.51	5.10	0.54	0.01	Test 4	5	2	11.07	4.97	0.55	0.00
Test 3	2	3	11.94	5.97	1.05	0.10	Test 4	2	3	11.82	5.81	0.99	0.11
Test 3	3	3	12.35	6.09	0.86	0.01	Test 4	3	3	11.62	5.68	0.78	0.01
Test 3	4	3	12.59	6.20	1.03	0.06	Test 4	4	3	11.65	5.72	0.89	0.06
Test 3	5	3	11.83	5.95	1.24	0.08	Test 4	5	3	10.72	5.44	1.04	0.07
Test 3	2	4	10.99	5.49	1.11	0.11	Test 4	2	4	10.47	5.30	1.03	0.09
Test 3	3	4	11.19	5.80	1.11	0.11	Test 4	3	4	10.36	5.23	1.01	0.10
Test 3	4	4	11.46	5.17	1.12	0.03	Test 4	4	4	10.31	4.69	0.96	0.02
Test 3	5	4	11.42	5.37	1.06	0.11	Test 4	5	4	10.10	4.81	0.92	0.09
Test 3	2	5	11.30	5.66	1.25	0.11	Test 4	2	5	10.87	5.39	1.18	0.09
Test 3	3	5	11.13	5.51	1.01	0.08	Test 4	3	5	10.16	4.82	0.79	0.04
Test 3	4	5	11.65	5.95	1.23	0.12	Test 4	4	5	10.43	5.30	1.03	0.10
Test 3	5	5	10.84	5.38	1.26	0.12	Test 4	5	5	9.38	4.54	1.01	0.09
	rows	M	10%	5%	1%	0.1%		rows	M	10%	5%	1%	0.1%
Test 5	2	2	8.46	3.14	0.21	0.00	Test 6	2	2	9.51	3.79	0.32	0.00
Test 5	3	2	7.44	2.82	0.25	0.00	Test 6	3	2	7.56	2.98	0.28	0.00
Test 5	4	2	7.48	2.99	0.26	0.01	Test 6	4	2	7.00	2.77	0.26	0.00
Test 5	5	2	7.32	2.83	0.24	0.00	Test 6	5	2	6.73	2.54	0.21	0.00
Test 5	2	3	10.27	4.88	0.78	0.05	Test 6	2	3	10.05	4.74	0.70	0.09
Test 5	3	3	10.55	4.88	0.68	0.01	Test 6	3	3	9.32	4.19	0.53	0.01
Test 5	4	3	10.52	4.88	0.72	0.05	Test 6	4	3	8.71	4.03	0.58	0.03
Test 5	5	3	9.70	4.74	0.81	0.06	Test 6	5	3	8.03	3.70	0.63	0.03
Test 5	2	4	10.01	5.06	0.90	0.08	Test 6	2	4	9.17	4.60	0.76	0.07
Test 5	3	4	10.16	5.14	0.95	0.10	Test 6	3	4	8.70	4.27	0.79	0.08
Test 5	4	4	10.30	4.68	0.93	0.01	Test 6	4	4	8.13	3.60	0.60	0.00
Test 5	5	4	10.11	4.75	0.93	0.09	Test 6	5	4	7.64	3.52	0.66	0.05
Test 5	2	5	10.89	5.40	1.16	0.09	Test 6	2	5	9.80	4.80	1.00	0.06
Test 5	3	5	10.50	5.13	0.88	0.05	Test 6	3	5	8.78	3.98	0.55	0.02
Test 5	4	5	10.90	5.59	1.10	0.11	Test 6	4	5	8.66	4.16	0.77	0.06
Test 5	5	5	10.02	4.92	1.12	0.10	Test 6	5	5	7.35	3.54	0.74	0.03

A Weighted Moving Average Process for Forecasting

Shou Hsing Shih Chris P. Tsokos
University of South Florida

The object of the present study is to propose a forecasting model for a nonstationary stochastic realization. The subject model is based on modifying a given time series into a new k-time moving average time series to begin the development of the model. The study is based on the autoregressive integrated moving average process along with its analytical constraints. The analytical procedure of the proposed model is given. A stock XYZ selected from the Fortune 500 list of companies and its daily closing price constitute the time series. Both the classical and proposed forecasting models were developed and a comparison of the accuracy of their responses is given.

Key words: ARIMA, moving average, stock, time series analysis

Introduction

Time series analysis and modeling plays a very important role in forecasting, especially when our initial stochastic realization is nonstationary in nature. Some of the interesting and useful publications related to the subject area are Akaike (1974), Banerjee et al. (1993), Box et al. (1994), Brockwell and Davis (1996), Dickey and Fuller (1979), Dickey et al. (1984), Durbin and Koopman (2001), Gardner et al. (1980), Harvey (1993), Jones (1980), Kwiatkowski et al. (1992), Rogers (1986), Said and Dickey (1984), Sakamoto et al. (1986), Shumway and Stoffer (2006), Tsokos (1973), Wei (2006).

The subject of the present study is to begin with a given time series that characterizes

an economic or any other natural phenomenon and as usual, is nonstationary. Box and Jenkins (1994) have introduced a popular and useful classical procedure to develop forecasting models that have been shown to be quite effective. In the present study, we introduce a procedure for developing a forecasting model that is more effective than the classical approach is introduced. For a given stationary or nonstationary time series, $\{x_t\}$, generate a k-day moving average time series, $\{y_t\}$, and the developmental process begins.

Basic concepts and analytical methods are reviewed that are essential in structuring the proposed forecasting model. The review is based on the autoregressive integrated moving average processes. The accuracy of the proposed forecasting model is illustrated by selecting from the list of Fortune 500 companies, company XYZ, and considering its daily closing prices for 500 days. The classical time series model for the subject information along with the proposed process was developed. A statistical comparison based on the actual and forecasting residuals is given, both in tabular and graphical form.

Proposed Forecasting Model: k-th Moving Average

Before introducing the proposed forecasting model, several important mathematical concepts will be defined that are essential in developing the analytical process. It

Shou Hsing Shih's research interests are in developing forecasting and statistical analysis and modeling of economic and environmental problems. Email: sshih@mail.usf.edu. Chris P. Tsokos is a Distinguished University Professor in mathematics and Statistics. His research interests are in modeling global warming; analysis and modeling of cancer data; parametric, Bayesian and nonparametric reliability; and stochastic systems. He is a fellow of ASA and ISI. Email: profcpt@cas.usf.edu

A WEIGHTED MOVING AVERAGE PROCESS FOR FORECASTING

is known that it is not possible to proceed in building a time series model without conforming to certain mathematical constraints such as stationarity of a given stochastic realization. Almost always, the time series that are given are nonstationary in nature and then, it is necessary to proceed to reduce it into being stationary. Let $\{x_t\}$ be the original time series. The difference filter is given by

$$(1 - B)^d, \quad (1)$$

where $B^j x_t = x_{t-j}$, and d is the degree of differencing of the series.

In time series analysis, the primary use for the k -th moving average process is for smoothing a realized time series. It is very useful in discovering a short-term, long-term trends and seasonal components of a given time series. The k -th moving average process of a time series $\{x_t\}$ is defined as follows:

$$y_t = \frac{1}{k} \sum_{j=0}^{k-1} x_{t-k+1+j}, \quad (2)$$

where $t = k, k+1, \dots, n$. It can be seen that as k increases, the number of observations k of $\{y_t\}$ decreases, and $\{y_t\}$ gets closer and closer to the mean of $\{x_t\}$ as k increases. In addition, when $k = n$, $\{y_t\}$ reduces to only a single observation, and equals μ , that is

$$y_t = \frac{1}{n} \sum_{j=1}^n x_j = \mu, \quad (3)$$

The proposed model is developed by transforming the original time series $\{x_t\}$ into $\{y_t\}$ by applying (2). After establishing the new time series, usually nonstationary, the process of reducing it into a stationary time series is begun. Kwiatkowski, Phillips, Schmit, and Shin (1992) introduced the KPSS Test to check the level of stationarity of a time series. The differencing order d is applied to the new time series $\{y_t\}$ for $d = 0, 1, 2, \dots$, then verify the stationarity of the series with the KPSS test until the series become stationary. Therefore, the

nonstationary time series is reduced into a stationary one after a proper number of differencing. The model building procedure is then developed via the proposed forecasting model.

After choosing a proper degree of differencing d , assume different orders for the autoregressive integrated moving average model, ARIMA(p, d, q), also known as Box and Jenkins method, where (p, d, q) represent the order of the autoregressive process, the order of differencing and the order of the moving average process, respectively. The ARIMA(p, d, q) is defined as follows:

$$\phi_p(B)(1 - B)^d y_t = \theta_q(B)\varepsilon_t, \quad (4)$$

where $\{y_t\}$ is the realized time series, ϕ_p and θ_q are the weights or coefficients of the AR and MA that drive the model, respectively, and ε_t is the random error. Write ϕ_p and θ_q as

$$\phi_p(B) = (1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p), \quad (5)$$

and

$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q). \quad (6)$$

In time series analysis, sometimes it is very difficult to make a decision in selecting the best order of the ARIMA(p, d, q) model when there are several models that all adequately represent a given set of time series. Hence, Akaike's information criterion (AIC) (1974), plays a major role when it comes to model selection. AIC was introduced by Akaike in 1973, and it is defined as:

$$AIC(M) = -2\ln[\text{maximum likelihood}] + 2M, \quad (7)$$

where M is the number of parameters in the model and the unconditional log-likelihood function suggested by Box, Jenkins, and Reinsel (1994), is given by

$$\ln L(\phi, \mu, \theta, \sigma_\varepsilon^2) = -\frac{n}{2} \ln 2\pi\sigma_\varepsilon^2 - \frac{S(\phi, \mu, \theta)}{2\sigma_\varepsilon^2}, \quad (8)$$

where $S(\phi, \mu, \theta)$ is the unconditional sum of squares function given by

$$S(\phi, \mu, \theta) = \sum_{t=-\infty}^n [E(\varepsilon_t | \phi, \mu, \theta, y)]^2 \quad (9)$$

where $E(\varepsilon_t | \phi, \mu, \theta, y)$ is the conditional expectation of ε_t given ϕ, μ, θ, y .

The quantities $\hat{\phi}$, $\hat{\mu}$, and $\hat{\theta}$ that maximize (8) are called unconditional maximum likelihood estimators. Since $\ln L(\phi, \mu, \theta, \sigma_\varepsilon^2)$ involves the data only through $S(\phi, \mu, \theta)$, these unconditional maximum likelihood estimators are equivalent to the unconditional least squares estimators obtained by minimizing $S(\phi, \mu, \theta)$. In practice, the summation in (9) is approximated by a finite form

$$S(\phi, \mu, \theta) = \sum_{t=M}^n [E(\varepsilon_t | \phi, \mu, \theta, y)]^2 \quad (10)$$

where M is a sufficiently large integer such that the back cast increment $|E(\varepsilon_t | \phi, \mu, \theta, y) - E(\varepsilon_{t-1} | \phi, \mu, \theta, y)|$ is less than any arbitrary predetermined small ε value for $t \leq -(M+1)$. This expression implies that $E(\varepsilon_t | \phi, \mu, \theta, y) \equiv \mu$; hence, $E(\varepsilon_t | \phi, \mu, \theta, y)$ is negligible for $t \leq -(M+1)$.

After obtaining the parameter estimates $\hat{\phi}$, $\hat{\mu}$, and $\hat{\theta}$, the estimate $\hat{\sigma}_\varepsilon^2$ of σ_ε^2 can then be calculated from

$$\hat{\sigma}_\varepsilon^2 = \frac{S(\hat{\phi}, \hat{\mu}, \hat{\theta})}{n} \quad (11)$$

For an ARMA(p,q) model based on n observations, the log-likelihood function is

$$\ln L = -\frac{n}{2} \ln 2\pi\sigma_\varepsilon^2 - \frac{1}{2\sigma_\varepsilon^2} S(\phi, \mu, \theta) \quad (12)$$

Proceed to maximize (12) with respect to the parameters ϕ, μ, θ , and σ_ε^2 , from (11),

$$\ln \hat{L} = -\frac{n}{2} \ln \hat{\sigma}_\varepsilon^2 - \frac{n}{2} (1 + \ln 2\pi) \quad (13)$$

Because the second term in expression (13) is a constant, we can reduce the AIC to the following expression

$$\text{AIC}(M) = n \ln \hat{\sigma}_\varepsilon^2 + 2M \quad (14)$$

Then, an appropriate time series model is generated and the statistical process with the smallest AIC can be selected. The model identified will possess the smallest average mean square error. The development of the model is summarized as follows.

- Transform the original time series $\{x_t\}$ into a new series $\{y_t\}$.
- Check for stationarity of the new time series $\{y_t\}$ by determining the order of differencing d , where $d = 0, 1, 2, \dots$ according to KPSS test, until stationarity is achieved.
- Decide the order m of the process. For this case, let $m = 5$ where $p + q = m$.
- After (d, m) is selected, list all possible set of (p, q) for $p + q \leq m$.
- For each set of (p, q) , estimate the parameters of each model, that is, $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$.
- Compute the AIC for each model, and choose the one with smallest AIC.

According to the criterion mentioned above, the ARIMA(p,d,q) model can be obtained that best fit a given time series, where the coefficients are $\phi_1, \phi_2, \dots, \phi_p, \theta_1, \theta_2, \dots, \theta_q$.

Using the model that we developed for $\{y_t\}$ and subject to the AIC criteria, we forecast values of $\{y_t\}$ and proceed to apply the back-shift operator to obtain estimates of the original phenomenon $\{x_t\}$, that is,

$$\hat{x}_t = k \hat{y}_t - x_{t-1} - x_{t-2} - \dots - x_{t-k+1} \quad (15)$$

The proposed model and the corresponding procedure discussed in this section shall be illustrated with real economic application and the results will be compared with the classical time series model.

A WEIGHTED MOVING AVERAGE PROCESS FOR FORECASTING

Figure 1. Daily Closing Price for Stock XYZ

Figure 2. Comparisons on Classical ARIMA Model VS. Original Time Series for the Last 100 Observations

Figure 3. Time Series Plot of the Residuals for Classical Model

Table 1. Basic Evaluation Statistics

$\bar{r}$	S_r^2	S_r	$\frac{S_r}{\sqrt{n}}$
0.02209169	0.1445187	0.3801562	0.0170011

Application: Forecasting Stock XYZ

A stock was selected from Fortune 500 companies that we identify as (XYZ). The daily closing price for 500 days constitutes the time series $\{x_t\}$. A plot of the actual information is given by Figure 1.

First, develop a time series forecasting model of the given nonstationary data using the ordinary Box and Jenkins methodology. Secondly, we shall modify the given data, Figure 1, to develop the proposed time series forecasting model. A comparison of the two models will be given.

The general theoretical form of the ARIMA(p,d,q) is given by

$$\phi_p(B)(1-B)^d x_t = \theta_q(B)\varepsilon_t \quad (16)$$

Following the Box and Jenkins' methodology (1994), the classical forecasting model with the

best AIC score is the ARIMA(1,1,2). That is, a combination of first order autoregressive (AR) and a second order moving average (MA) with a first difference filter. Thus, write it as

$$(1 - .9631B)(1 - B)x_t = (1 - 1.0531B + .0581B^2)\varepsilon_t \quad (17)$$

After expanding the autoregressive operator and the difference filter,

$$(1 - 1.9631B + .9631B^2)x_t = (1 - 1.0531B + .0581B^2)\varepsilon_t \quad (18)$$

and rewrite the model as

A WEIGHTED MOVING AVERAGE PROCESS FOR FORECASTING

Table 2. Actual and Predicted Price

N	Actual Price	Predicted Price	Residuals
476	26.78	26.8473	-0.0673
477	26.75	26.7976	-0.0476
478	26.67	26.7673	-0.0972
479	26.8	26.6922	0.1078
480	26.73	26.8064	-0.0764
481	26.78	26.7490	0.0310
482	26.27	26.7911	-0.5211
483	26.12	26.3277	-0.2077
484	26.32	26.1631	0.1569
485	25.98	26.3364	-0.3564
486	25.86	26.0349	-0.1749
487	25.65	25.9068	-0.2568
488	25.67	25.6670	0.0031
489	26.02	25.7119	0.3081
490	26.01	26.0335	-0.0235
491	26.11	26.0427	0.0674
492	26.18	26.1343	0.0457
493	26.28	26.2032	0.0768
494	26.39	26.2986	0.0914
495	26.46	26.4043	0.0557
496	26.18	26.4743	-0.2943
497	26.32	26.2219	0.0981
498	26.16	26.3354	-0.1754
499	26.24	26.1953	0.0447
500	26.07	26.2602	-0.1902

$$x_t = 1.9631x_{t-1} - .9631x_{t-2} + \varepsilon_t - 1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2} \quad (19)$$

by letting $\varepsilon_t = 0$, there is the one day ahead forecasting time series of the closing price of stock XYZ as

$$\hat{x}_t = 1.9631x_{t-1} - .9631x_{t-2} - 1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2}. \quad (20)$$

Using the above equation, graph the forecasting values obtained by using the classical approach on top of the original time series, as shown by Figure 2.

The basic statistics that reflect the accuracy of model (20) are the mean $\bar{r}$, variance S_r^2 , standard deviation S_r , and standard error $S_r/\sqrt{n}$ of the residuals. Figure 3 gives a plot of the residual and Table 1 gives the basic statistics.

Furthermore, restructure the model (20) with $n = 475$ data points to forecast the last 25

Figure 4. Three Days Moving Average on Daily Closing Price of Stock XYZ VS. the Original Time Series

observations using only the previous information. The purpose is to see how accurate our forecast prices are with respect to the actual 25 values that have not been used. Table 2 gives the actual price, predicted price, and residuals between the forecasts and the 25 hidden values.

The average of these residuals is $\bar{r} = -0.05608$.

Proceed to develop the proposed forecasting model. The original time series of stock XYZ daily closing prices is given by Figure 1. The

new time series is being created by $k = 3$ days moving average and the analytical form of $\{y_t\}$ is given by

$$y_t = \frac{x_{t-2} + x_{t-1} + x_t}{3} \quad (21)$$

Figure 4 shows the new time series $\{y_t\}$ along with the original time series $\{x_t\}$, that will be used to develop the proposed forecasting model.

Following the procedure stated above, the best model that characterizes the behavior of $\{y_t\}$ is ARIMA (2,1,3). That is,

$$(1 - .8961B - .0605B^2)(1 - B)y_t = (1 + .0056B - .0056B^2 - B^3)\varepsilon_t \quad (22)$$

Expanding the autoregressive operator and the first difference filter, we have

$$(1 - 1.8961B + .8356B^2 + .0605B^3)y_t = (1 + .0056B - .0056B^2 - B^3)\varepsilon_t \quad (23)$$

Thus, write (23) as

$$y_t = 1.8961y_{t-1} - .8356y_{t-2} - .0605y_{t-3} + \varepsilon_t + .0056\varepsilon_{t-1} - .0056\varepsilon_{t-2} - \varepsilon_{t-3} \quad (24)$$

The final analytical form of the proposed forecasting model can be written as

$$\hat{y}_t = 1.8961y_{t-1} - .8356y_{t-2} - .0605y_{t-3} + .0056\varepsilon_{t-1} - .0056\varepsilon_{t-2} - \varepsilon_{t-3} \quad (25)$$

Using the above equation, a plot of the developed model (25), showing a one day ahead

A WEIGHTED MOVING AVERAGE PROCESS FOR FORECASTING

Figure 5. Comparisons on Our Proposed Model VS. Original Time Series for the Last 100 Observations

Figure 6. Time Series Plot for Residuals for Our Proposed Model

Table 3. Basic Evaluation Statistics

$\bar{r}$	S_r^2	S_r	$\frac{S_r}{\sqrt{n}}$
0.01016814	0.1437259	0.3791119	0.01698841

Table 4. Actual and Predicted Price

N	Actual Price	Predicted Price	Residuals
476	26.78	26.8931	-0.1131
477	26.75	26.7715	-0.0215
478	26.67	26.7121	-0.0421
479	26.8	26.7239	0.0761
480	26.73	26.7854	-0.0554
481	26.78	26.6892	0.0908
482	26.27	26.8292	-0.5592
483	26.12	26.3027	-0.1827
484	26.32	26.0808	0.2392
485	25.98	26.3603	-0.3803
486	25.86	25.9868	-0.1268
487	25.65	25.8443	-0.1943
488	25.67	25.7115	-0.0414
489	26.02	25.6499	0.3701
490	26.01	25.9650	0.0450
491	26.11	26.0526	0.0574
492	26.18	26.0912	0.0888
493	26.28	26.1449	0.1351
494	26.39	26.3090	0.0810
495	26.46	26.3752	0.0848
496	26.18	26.4223	-0.2423
497	26.32	26.2461	0.0739
498	26.16	26.2964	-0.1364
499	26.24	26.1437	0.0963
500	26.07	26.2678	-0.1978

A WEIGHTED MOVING AVERAGE PROCESS FOR FORECASTING

Table 5. Basic Comparison on Classical Approach VS. Our Proposed Model

	$\bar{r}$	S_r^2	S_r	$\frac{S_r}{\sqrt{n}}$
Classical	0.02209169	0.1445187	0.3801562	0.0170011
Proposed	0.01016814	0.1437259	0.3791119	0.01698841

forecasting along with the new time series, $\{y_t\}$, is displayed by Figure 5. Note the closeness of the two plots that reflect the quality of the proposed model.

Similar to the classical model approach that we discussed earlier, use the first 475 observations $\{y_1, y_2, \dots, y_{475}\}$ to forecast $\hat{y}_{476}$. Then, use the observations $\{y_1, y_2, \dots, y_{476}\}$ to forecast $\hat{y}_{477}$, and continue this process until forecasts are obtained for all the observations, that is, $\{\hat{y}_{476}, \hat{y}_{477}, \dots, \hat{y}_{500}\}$. From equation (21), the relationship can be seen between the forecasting values of the original series $\{x_t\}$ and the forecasting values of 3 days moving average series $\{y_t\}$, that is,

$$\hat{x}_t = 3\hat{y}_t - x_{t-1} - x_{t-2} \quad (26)$$

Hence, after $\{\hat{y}_{476}, \hat{y}_{477}, \dots, \hat{y}_{500}\}$ is estimated, use the above equation, (26), to solve the forecasting values for $\{x_t\}$. Figure 6 is the residual plot generated by the proposed model, and followed by Table 3, that includes the basic evaluation statistics.

Both of the above displayed evaluations reflect on accuracy of the proposed model. The actual daily closing prices of stock XYZ from the 476th day along with the forecasted prices and residuals are given in Table 4.

The results given above attest to the good forecasting estimates for the hidden data.

Comparison of the Forecasting Models

The two developed models are compared. The classical process is given by

$$\hat{x}_t = 1.9631x_{t-1} - .9631x_{t-2} - 1.0531\varepsilon_{t-1} + .0581\varepsilon_{t-2} \quad (27)$$

In the proposed model, the following inversion is used to obtain the estimated daily closing prices of stock XYZ, that is,

$$\hat{y}_t = 1.8961y_{t-1} - .8356y_{t-2} - .0605y_{t-3} + .0056\varepsilon_{t-1} - .0056\varepsilon_{t-2} - \varepsilon_{t-3} \quad (28)$$

in conjunction with

$$\hat{x}_t = 3\hat{y}_t - x_{t-1} - x_{t-2} \quad (29)$$

Table 5 given is a comparison of the basic statistics used to evaluate the two models under investigation. The average mean residuals between the two models show that the proposed model is overall approximately 54% more effective in estimating one day ahead the closing price of Fortune 500 stock XYZ.

Conclusion

In the present study a new time series model is introduced that is based on the actual stochastic realization of a given phenomenon. The proposed model is based on modifying the given economic time series, $\{x_t\}$, and smoothing it with k-time moving average to create a new time

series $\{y_t\}$. The basic analytical procedures are developed through the developing process of a forecasting model. A step-by-step procedure is memorized for the final computational procedure for a nonstationary time series. To evaluate the effectiveness of our proposed model We selected a company from the Fortune 500 list, company XYZ the daily closing prices of the stock for 500 days was used as our time series data , $\{x_t\}$, which was as usual nonstationary. We developed the classical time series forecasting model using the Box and Jenkins methodology and also our proposed model, $\{y_t\}$, based on a 3-way moving average smoothing procedure. The analytical form of the two forecasting models is presented and a comparison of them is also given. Based on the average mean residuals, the proposed model was significantly more effective in such terms of predicting of the closing daily prices of the stock XYZ.

References

- Akaike, H. (1974). A New Look at the Statistical Model Identification, *IEEE Transactions on Automatic Control*, AC-19, 716-723.
- Banerjee, A., Dolado, J. J., Galbraith, J. W., & Hendry, D. F. (1993). *Cointegration, Error Correction, and the Econometric Analysis of Non-Stationary Data*, Oxford University Press, Oxford.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ., 89-99.
- Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994) *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ., 224-247.
- Brockwell, P. J., & Davis, R. A. (1996). *Introduction to Time Series and Forecasting.*, Springer, New York., Sections 3.3 and 8.3.
- Dickey, D. A., & Fuller, W. A. (1979) Distribution and the Estimators for Autoregressive Time Series With a Unit Root., *Journal of the American Statistical Association*, Vol. 74, No. 366, 427-431.
- Dickey, D. A., Hasza, D. P., & Fuller, W. A. (1984). Testing for Unit Roots in Seasonal Time Series., *Journal of the American Statistical Association*, Vol. 79, No. 386, 355-367.
- Durbin, J., & Koopman, S. J. (2001). *Time Series Analysis by State Space Methods.*, Oxford University Press.
- Gardner, G., Harvey, A. C., & Phillips, G. D. A. (1980). Algorithm AS154. An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering., *Applied Statistics*, 29, 311-322.
- Harvey, A. C. (1993). *Time Series Models*, 2nd Edition, Harvester Wheatsheaf., sections 3.3 and 4.4.
- Jones, R. H. (1980). Maximum likelihood fitting of ARMA models to time series with missing observations., *Technometrics*, 20, 389-395.
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root., *Journal of Econometrics*, 54, 159-178.
- Rogers, A. J. (1986). Modified Lagrange Multiplier Tests for Problems with One-Sided Alternatives, *Journal of Econometrics*, North-Holland., 31, 341-361.
- Said, S. E., & Dickey, D. A. (1984) Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order., *Biometrika*, 71, 599-607.
- Sakamoto, Y., Ishiguro, M., & Kitagawa, G. (1986). *Akaike Information Criterion Statistics.*, D. Reidel Publishing Company.
- Shumway, R. H., & Stoffer, D. S. (2006). *Time Series Analysis and Its Applications: with R Examples*, 2nd ed., Springer, New York.
- Tsokos, C. P. (1973). Forecasting Models from Non-Stationary Time Series-Short Term Predictability of Stocks., *Mathematical Methods in Investment and Finance.*, North Holland Publishing Co., 520-63.
- Wei, W. W. S. (2006). *Time Series Analysis: Univariate and Multivariate Methods*, 2nd ed., Pearson Education, Inc.

Estimation of Covariance Matrix in Signal Processing When the Noise Covariance Matrix is Arbitrary

Madhusudan Bhandary
Columbus State University

An estimator of the covariance matrix in signal processing is derived when the noise covariance matrix is arbitrary based on the method of maximum likelihood estimation. The estimator is a continuous function of the eigenvalues and eigenvectors of the matrix $\hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}}$, where S^* is the sample covariance matrix of observations consisting of both noise and signals and $\hat{\Sigma}_1$ is the estimator of covariance matrix based on observations consisting of noise only. Strong consistency and asymptotic normality of the estimator are briefly discussed.

Key words: Maximum likelihood estimator, signal processing, white noise, colored noise.

Introduction

The covariance and correlation matrices are used for a variety of purposes. They give a simple description of the overall shape of a point-cloud in p-space. They are used in principal component analysis, factor analysis, discriminant analysis, canonical correlation analysis, tests of independence etc. In signal processing, estimation of covariance matrix is important because it helps to discriminate between signals and noise (filtering).

The problem of estimation of the dispersion matrix of the form $\Gamma + \sigma^2 \Sigma_1$ is considered, where the unknown matrix Γ is n.n.d. of rank $q(< p)$, $\sigma^2(> 0)$ is unknown and Σ_1 is some arbitrary positive matrix. In general, the model is signal processing is

$$\mathbf{X}(t) = \mathbf{A}\mathbf{S}(t) + \mathbf{n}(t) \quad (1.1)$$

where, $\mathbf{X}(t) = (X_1(t), X_2(t), \dots, X_p(t))'$ is the $p \times 1$ observation vector at time t , $\mathbf{S}(t) = (S_1(t), S_2(t), \dots, S_q(t))'$ is the $q \times 1$ vector of unknown random signals at time t , $\mathbf{n}(t) = (n_1(t), n_2(t), \dots, n_p(t))'$ is the $p \times 1$ random noise vector at time t , and $\mathbf{A} = (\mathbf{A}(\Phi_1), \mathbf{A}(\Phi_2), \dots, \mathbf{A}(\Phi_q))$ is the $p \times q$ matrix of unknown coefficients, $\mathbf{A}(\Phi_r)$ is the $p \times 1$ vector of functions of the elements of unknown vector Φ_r associated with the r^{th} signal and $q < p$.

In model (1.1), $\mathbf{X}(t)$ is assumed to be distributed as p -variate normal distribution with mean vector zero and dispersion matrix $\mathbf{A}\Psi\mathbf{A}' + \sigma^2 \Sigma_1 = \Gamma + \sigma^2 \Sigma_1$, where $\Gamma = \mathbf{A}\Psi\mathbf{A}'$ is unknown n.n.d. matrix of rank $q(< p)$ and Ψ = covariance matrix of $\mathbf{S}(t)$, $\sigma^2(> 0)$ is unknown, $\sigma^2 \Sigma_1$ is the covariance matrix of the noise vector $\mathbf{n}(t)$ and Σ_1 is some arbitrary positive definite matrix. In the above situation, when the covariance matrix of the noise vector $\mathbf{n}(t)$ is $\sigma^2 I_p$, where I_p denotes identity matrix of order $p \times p$, the model is called white noise model. If the covariance matrix of $\mathbf{n}(t)$ is $\sigma^2 \Sigma_1$, where Σ_1 is some arbitrary positive definite matrix, the model is colored noise model.

One of the important problems that arise in the area of signal processing is to estimate q , the number of signals transmitted. The problem

Madhusudan Bhandary is on the faculty of the Department of Mathematics. Mailing address: Columbus State University, 4225 University Avenue, Columbus, GA 31907. E-mail: bhandary_madhusudan@colstate.edu.

is equivalent to estimate the multiplicity of the smallest eigen value of the covariance matrix of the observation vector. Anderson (1963), Krishnaiah (1976), Rao (1983), Wax and Kailath (1984), Zhao et.al (1986a,b) considered the above problem. Chen (2001), Chen (2002) and Kundu (2000) developed procedures for estimating the number of signals.

Another important problem in this area is to have some idea about covariance and correlation matrix. The estimation of the dispersion matrix of the form $\Gamma + \sigma^2 \Sigma_1$ is of interest, and then, the derivation of the estimator is discussed. Strong consistency and asymptotic normality of the estimator are then discussed.

Derivation of the Estimator

Let the observations $\mathbf{x}(t_1), \mathbf{x}(t_2), \dots, \mathbf{x}(t_n)$ be n observed p -component signals at n different time points which are independently and identically distributed as p -variate normal distribution with mean vector zero and dispersion matrix $\Gamma + \sigma^2 \Sigma_1$, where $\Gamma = A\Psi A'$ and is n.n.d. of rank $q(<p)$ and Σ_1 is some arbitrary positive definite matrix.

Because Γ is n.n.d. of rank $q(<p)$, it can be assumed that $\Gamma = BB'$, where B is a pxq matrix of rank q and

$$B'B = \text{Diag}(\theta_1, \theta_2, \dots, \theta_q), \quad (2.1)$$

where $\theta_1 \geq \theta_2 \geq \dots \geq \theta_q$ are the non-zero eigen values of Γ .

The log-likelihood of the observations based on $\mathbf{x}_i$'s, apart from a constant term, can be written as follows :

$$\log L = -\frac{n}{2} \log |BB' + \sigma^2 \Sigma_1| - \frac{1}{2} \text{tr}((BB' + \sigma^2 \Sigma_1)^{-1} S) \quad (2.2)$$

where, $S = \sum_{i=1}^n x_i x_i' = \sum_{i=1}^n x(t_i) x(t_i)', i=1,2,\dots,n$

Following Lawley and Maxel (1963, Chapter 2):

$$\begin{aligned} \frac{\partial \log L}{\partial B} &= \left[-\frac{n}{2} (BB' + \sigma^2 \Sigma_1)^{-1} + \frac{1}{2} (BB' + \sigma^2 \Sigma_1)^{-1} S (BB' + \sigma^2 \Sigma_1)^{-1} \right] \\ 2B &= 0 \\ \text{i.e. } \Sigma_2^{-1} (\Sigma_2 - S^*) \Sigma_2^{-1} B &= 0 \quad (2.3) \end{aligned}$$

where, $\Sigma_2 = BB' + \sigma^2 \Sigma_1$ and $S^* = \frac{S}{n}$.

Using Rao(1983, p.33)

$$\Sigma_2^{-1} = (BB' + \sigma^2 \Sigma_1)^{-1} =$$

$$\begin{aligned} & \left(\frac{\Sigma_1^{-1}}{\sigma^2} - \frac{\Sigma_1^{-1}}{\sigma^2} B \left(\frac{B' \Sigma_1^{-1} B}{\sigma^2} + I_q \right)^{-1} \frac{B' \Sigma_1^{-1}}{\sigma^2} \right) = \\ & \frac{1}{\sigma^2} (\Sigma_1^{-1} - \Sigma_1^{-1} B (I_q + D)^{-1} \frac{B' \Sigma_1^{-1}}{\sigma^2}) \quad (2.4) \end{aligned}$$

where, $D = \frac{B' \Sigma_1^{-1} B}{\sigma^2}$ and I_p denotes identity matrix of order pxp . Using (2.4) in (2.3),

$$\begin{aligned} & (\Sigma_2 - S^*) \frac{1}{\sigma^2} \left[\Sigma_1^{-1} - \Sigma_1^{-1} B (I_q + D)^{-1} \frac{B' \Sigma_1^{-1}}{\sigma^2} \right] B = 0 \\ \text{i.e. } & (\Sigma_2 - S^*) \frac{\Sigma_1^{-1} B}{\sigma^2} [I_q - (I_q + D)^{-1} D] = 0 \\ \text{i.e. } & (\Sigma_2 - S^*) \frac{\Sigma_1^{-1} B}{\sigma^2} (I_q + D)^{-1} = 0 \\ \text{i.e. } & (\Sigma_2 - S^*) \Sigma_1^{-1} B = 0 \quad (2.5) \end{aligned}$$

which after substitution of Σ_2 from (2.3) and rearrangement of terms gives

$$\begin{aligned} & S^* \Sigma_1^{-1} B = B (\sigma^2 I_q + B' \Sigma_1^{-1} B) \\ \text{i.e. } & (\Sigma_1^{-\frac{1}{2}} S^* \Sigma_1^{-\frac{1}{2}}) (\Sigma_1^{-\frac{1}{2}} B) = \\ & (\Sigma_1^{-\frac{1}{2}} B) (\sigma^2 I_q + B' \Sigma_1^{-1} B) \quad (2.6) \end{aligned}$$

It can be seen that the right hand side of (2.2) remains the same, if the matrix B is replaced by BP where P is an orthogonal matrix and hence $B'\Sigma_1^{-1}B$ can be reduced to $P'B'\Sigma_1^{-1}BP$ which can be reduced to a diagonal form because $B'\Sigma_1^{-1}B$ is a real symmetric matrix (See Bellman (1960) p.54).

From (2.6) it is trivial that columns of $\Sigma_1^{-\frac{1}{2}}B$ are eigenvectors of the matrix $\Sigma_1^{-\frac{1}{2}}S^*\Sigma_1^{-\frac{1}{2}}$ and the diagonal elements of $\sigma^2 I_q + B'\Sigma_1^{-1}B$ are the corresponding eigenvalues (2.7).

Let $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_p$ be the ordered eigen values of $\Sigma_1^{-\frac{1}{2}}S^*\Sigma_1^{-\frac{1}{2}}$ and let $\Theta = \text{Diag.}(\alpha_1, \alpha_2, \dots, \alpha_q)$. Since the diagonal elements of $B'\Sigma_1^{-1}B$ are the column sum of squares of $\Sigma_1^{-\frac{1}{2}}B$, each eigenvector should be normalized so that the sum of squares equal the corresponding eigenvalue minus σ^2 . Let $\tilde{B}$ be a $p \times q$ matrix whose columns are $w_1, w_2, \dots, w_q$, where $w_1, w_2, \dots, w_q$ are a set of unit-length eigen vectors corresponding to the q largest eigen values of $\Sigma_1^{-\frac{1}{2}}S^*\Sigma_1^{-\frac{1}{2}}$. Then,

$$\tilde{B}'\tilde{B} = I_q$$

and

$$\Sigma_1^{-\frac{1}{2}}\hat{B} = \tilde{B}(\Theta - \sigma^2 I_q)^{\frac{1}{2}} \quad (2.8)$$

Another likelihood equation can be written as follows:

$$\frac{\partial \log L}{\partial \sigma^2} =$$

$$\text{tr.}(\Sigma_2^{-1}(\Sigma_2 - S^*)\Sigma_2^{-1}\Sigma_1) = 0 \quad (2.9)$$

From (2.4) and (2.9),

$$\text{tr.} \left[(I_p - \Sigma_2^{-1}S^*) \left(\frac{1}{\sigma^2} (I_p - \Sigma_1^{-1}B(I_q + D)^{-1} \frac{B'}{\sigma^2}) \right) \right] = 0,$$

$$\text{tr.} \left[\frac{1}{\sigma^2} (I_p - \Sigma_1^{-1}B(I_q + D)^{-1} \frac{B'}{\sigma^2}) - \frac{\Sigma_2^{-1}S^*}{\sigma^2} + \frac{1}{\sigma^2} \Sigma_2^{-1}S^*\Sigma_1^{-1}B(I_q + D)^{-1} \frac{B'}{\sigma^2} \right] = 0$$

$$\text{tr.} \left[\frac{I_p}{\sigma^2} - \frac{\Sigma_1^{-1}B}{\sigma^2} (I_q + D)^{-1} \frac{B'}{\sigma^2} - \frac{\Sigma_2^{-1}S^*}{\sigma^2} + \frac{\Sigma_1^{-1}B}{\sigma^2} (I_q + D)^{-1} \frac{B'}{\sigma^2} \right] = 0 \quad (\text{using (2.5)})$$

$$\text{tr.} \left[\frac{I_p}{\sigma^2} - \frac{\Sigma_2^{-1}S^*}{\sigma^2} \right] = 0$$

$$\text{tr.} \left[\frac{I_p}{\sigma^2} - \frac{1}{\sigma^2} \left\{ \Sigma_1^{-1} - \Sigma_1^{-1}B(I_q + D)^{-1} \frac{B'\Sigma_1^{-1}}{\sigma^2} \right\} \frac{S^*}{\sigma^2} \right] = 0 \quad (\text{using (2.4)})$$

i.e.,

$$\text{tr.} \left[\frac{I_p}{\sigma^2} - \frac{\Sigma_1^{-1}S^*}{\sigma^4} + \frac{\Sigma_1^{-1}B(I_q + D)^{-1}B'\Sigma_1^{-1}S^*}{\sigma^6} \right] = 0$$

$$\frac{p}{\sigma^2} - \frac{\text{tr.}(\Sigma_1^{-\frac{1}{2}}S^*\Sigma_1^{-\frac{1}{2}})}{\sigma^4} + \frac{\text{tr.}(B'\Sigma_1^{-1}B)}{\sigma^4} = 0 \quad (2.10)$$

(2.10) is obtained due to the fact that

$$\frac{\Sigma_1^{-1}B(I_q + D)^{-1}B'\Sigma_1^{-1}S^*}{\sigma^6} =$$

$$\frac{\Sigma_1^{-1}B(I_q + D)^{-1}B'\Sigma_1^{-1}\Sigma_2}{\sigma^6}$$

(using (2.5))

$$= \frac{\Sigma_1^{-1}B(I_q + D)^{-1}(I_q + D)\sigma^2 B'}{\sigma^6}$$

$$= \frac{\Sigma_1^{-1}BB'}{\sigma^4}$$

(because $B'\Sigma_1^{-1}\Sigma_2 = B'\Sigma_1^{-1}(BB' + \sigma^2 \Sigma_1)$)

$$= B'\Sigma_1^{-1}BB' + \sigma^2 B' =$$

$$(D + I_q)\sigma^2 B')$$

From (2.10),

$$\begin{aligned} & \frac{p}{\sigma^2} - \frac{\sum_{i=1}^p \alpha_i}{\sigma^4} + \frac{\text{tr}(\Theta - \sigma^2 I_q)}{\sigma^4} \\ & = 0 \text{ (using (2.8))} \\ \text{i.e. } & \frac{p}{\sigma^2} - \frac{\sum_{i=1}^p \alpha_i}{\sigma^4} + \frac{\sum_{i=1}^q (\alpha_i - \sigma^2)}{\sigma^4} = 0 \\ \text{i.e. } & \hat{\sigma}^2 = \frac{\sum_{i=q+1}^p \alpha_i}{p-q} \end{aligned} \quad (2.11)$$

It remains to estimate the matrix Σ_1 . An independent set of observations on noise is necessary to be found only to estimate Σ_1 . Let $y(t_1), y(t_2), \dots, y(t_m)$ be i.i.d. $\sim N_p(0, \sigma^2 \Sigma_1)$. Let $y(t_i) = y_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$ for convenience. Then the trivial estimator of the covariance matrix

$$\Sigma_1 \text{ is } \hat{\Sigma}_1 = \frac{1}{m} \sum_{i=1}^m y_i y_i' \quad (2.12)$$

Hence, final estimator of the covariance matrix can be written as follows:

$$\begin{aligned} \text{Estimator of } (\Gamma + \sigma^2 \Sigma_1) &= \hat{B} \hat{B}' + \hat{\sigma}^2 \hat{\Sigma}_1 \\ &= \hat{\Sigma}_1^{-\frac{1}{2}} \tilde{B} (\Theta - \hat{\sigma}^2 I_q) \tilde{B}' \hat{\Sigma}_1^{-\frac{1}{2}} + \hat{\sigma}^2 \hat{\Sigma}_1 \end{aligned} \quad (2.13)$$

where

$$\begin{aligned} \tilde{B} &= (w_1: w_2: \dots: w_q) \\ \Theta &= \text{Diag}(\alpha_1, \alpha_2, \dots, \alpha_q) \\ \alpha_r &= r^{\text{th}} \text{ ordered eigen value of } \hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}} \\ w_r &= r^{\text{th}} \text{ orthonormal eigenvector of } \\ & \quad \hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}} \\ & \quad \text{corresponding to } \alpha_r \\ \hat{\sigma}^2 & \text{ is given by (2.11)} \\ \text{and } \hat{\Sigma}_1 & \text{ can be obtained from (2.12).} \end{aligned}$$

Strong Consistency of the Estimator

Lemma 3.1.

Let the observations $y_1, y_2, \dots, y_m$ be i.i.d. $\sim N_p(0, \sigma^2 \Sigma_1)$, where Σ_1 is some arbitrary positive definite matrix. Let $\hat{\Sigma}_1$ be the estimator of Σ_1 given by (2.12). Then $\hat{\Sigma}_1$ is a strongly consistent estimator of Σ_1 .

Proof.

The proof of Lemma 3.1 is trivial from Strong Law of Large Number Theory.

Lemma 3.2

Suppose $A, A_n, n = 1, 2, \dots$, are all $p \times p$ symmetric matrices such that $A_n - A = O(\alpha_n)$ and $\alpha_n \rightarrow 0$ as $n \rightarrow \infty$. Denote by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ and $\lambda_1^{(n)} \geq \lambda_2^{(n)} \geq \dots \geq \lambda_p^{(n)}$ the eigenvalues of A and A_n , respectively. Then,

$$\lambda_i^{(n)} - \lambda_i = O(\alpha_n) \text{ as } n \rightarrow \infty, i = 1, \dots, p.$$

Proof.

The proof of Lemma 3.2 is given in Zhao, Krishnaiah and Bai (1986a).

Lemma 3.3

Suppose $A, A_n, n = 1, 2, \dots$, are all $p \times p$ symmetric matrices such that $A_n - A = O(\beta_n)$ and $\beta_n \rightarrow 0$ as $n \rightarrow \infty$. Denote $f_1, f_2, \dots, f_p$ and $f_1^{(n)}, f_2^{(n)}, \dots, f_p^{(n)}$ the eigenvectors of A and A_n respectively, corresponding to $\lambda_1, \lambda_2, \dots, \lambda_p$ and $\lambda_1^{(n)}, \lambda_2^{(n)}, \dots, \lambda_p^{(n)}$ respectively.

Then, $\|f_i^{(n)} - f_i\| = O(\beta_n)$ as $n \rightarrow \infty, i = 1, \dots, p$.

Note: Lemma 3.3 may not be true, if the symmetric matrix A has same eigenvalues. But it is true for those eigenvectors corresponding to distinct eigenvalues of A .

Proof.

The proof of Lemma 3.3 can be done similar way as in Zhao, Krishnaiah and Bai (1986a).

Theorem 3.1

Let $\Gamma + \hat{\sigma}^2 \Sigma_1$ be an estimator of $\Gamma + \sigma^2 \Sigma_1$ obtained from (2.13). Then $\Gamma + \hat{\sigma}^2 \Sigma_1 \xrightarrow{a.s.} \Gamma + \sigma^2 \Sigma_1$ as $n \rightarrow \infty$ and $m \rightarrow \infty$.

Proof.

Using Lemma 3.1,

$$\hat{\Sigma}_1 \xrightarrow{a.s.} \Sigma_1 \text{ as } m \rightarrow \infty \quad (3.1)$$

From Strong Law of Large Number Theory,

$$\begin{aligned} S^* &= \frac{1}{n} \sum_{i=1}^n x_i x_i' \xrightarrow{a.s.} E(x_i x_i') \\ &\text{as } n \rightarrow \infty \\ &= V(x_1) + 00' \\ &= \Gamma + \sigma^2 \Sigma_1 \end{aligned}$$

Hence,

$$\begin{aligned} \hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}} &\xrightarrow{a.s.} \Sigma_1^{-\frac{1}{2}} (\Gamma + \sigma^2 \Sigma_1) \Sigma_1^{-\frac{1}{2}} \\ &\text{as } n \rightarrow \infty \text{ and } m \rightarrow \infty \\ &= \Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p \end{aligned} \quad (3.2)$$

Let $l_1 > l_2 > \dots > l_q > \sigma^2$ be the ordered eigenvalues of $\Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p$ and $d_1, d_2, \dots, d_p$ be the corresponding orthonormal eigenvectors of $\Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p$. Then, using (3.2) and Lemma 3.2,

$$\alpha_i \xrightarrow{a.s.} l_i ; i = 1, 2, \dots, q$$

and

$$\alpha_i \xrightarrow{a.s.} \sigma^2 \text{ for } i = q+1, \dots, p \text{ as } n \rightarrow \infty \quad (3.3)$$

Because the eigenvalues $l_1, l_2, \dots, l_q$ of $\Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p$ are not the same, using (3.2) and Lemma 3.3,

$$\begin{aligned} w_i &\xrightarrow{a.s.} d_i ; i = 1, 2, \dots, q \\ &\text{as } n \rightarrow \infty \end{aligned} \quad (3.4)$$

where α_i 's and w_i 's are explained in (2.13).

$$\begin{aligned} \text{Now, } \hat{\sigma}^2 &= \frac{\sum_{i=q+1}^p \alpha_i}{p-q} \xrightarrow{a.s.} \sigma^2 \text{ as } n \rightarrow \infty \\ &\text{(using (3.3))} \end{aligned} \quad (3.5)$$

and

$$\begin{aligned} \Gamma + \hat{\sigma}^2 \Sigma_1 &= \hat{\Sigma}_1^{-\frac{1}{2}} \tilde{B} (\Theta - \hat{\sigma}^2 I_q) \tilde{B}' \hat{\Sigma}_1^{-\frac{1}{2}} + \hat{\sigma}^2 \hat{\Sigma}_1 \\ &= \hat{\Sigma}_1^{-\frac{1}{2}} \left(\sum_{i=1}^q (\alpha_i - \hat{\sigma}^2) w_i w_i' \right) \hat{\Sigma}_1^{-\frac{1}{2}} + \hat{\sigma}^2 \hat{\Sigma}_1 \\ &\xrightarrow{a.s.} \Sigma_1^{-\frac{1}{2}} \left(\sum_{i=1}^q (l_i - \sigma^2) d_i d_i' \right) \Sigma_1^{-\frac{1}{2}} + \sigma^2 \Sigma_1 \end{aligned} \quad (3.6)$$

Because $d_1, d_2, \dots, d_p$ are orthonormal eigenvectors,

$$DD' = I_p \text{ where } D = (d_1 : d_2 : \dots : d_p)_{p \times p}$$

Hence,

$$\sigma^2 I_p = \sigma^2 \sum_{i=1}^q d_i d_i' + \sigma^2 \sum_{i=q+1}^p d_i d_i' \quad (3.7)$$

Again, from Spectral Decomposition,

$$\begin{aligned} \Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p &= \\ \sum_{i=1}^q l_i d_i d_i' + \sigma^2 \sum_{i=q+1}^p d_i d_i' \end{aligned} \quad (3.8)$$

Therefore,

$$\begin{aligned}
 & \sum_{i=1}^q (l_i - \sigma^2) d_i d_i' \\
 &= \sum_{i=1}^q l_i d_i d_i' - \sigma^2 \sum_{i=1}^q d_i d_i' \\
 &= (\Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p - \sigma^2 \sum_{i=q+1}^p d_i d_i') - \\
 & \quad (\sigma^2 I_p - \sigma^2 \sum_{i=q+1}^p d_i d_i') \\
 & \quad (\text{ using (3.7) and (3.8) }) \\
 &= \Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} \quad (3.9)
 \end{aligned}$$

Using (3.9) in (3.6), we get Theorem 3.1.

Asymptotic Normality of the Estimator
Theorem 4.1

Let $\hat{\Gamma} + \hat{\sigma}^2 \Sigma_1$ be an estimator of $\Gamma + \sigma^2 \Sigma_1$ obtained from (2.13).

Then the limiting distribution of $\sqrt{n} (\hat{\Gamma} + \hat{\sigma}^2 \Sigma_1 - \Gamma + \sigma^2 \Sigma_1)$ is normal with mean 0 and variance B where B is given by (4.5) later.

Proof.

From (3.1) $\hat{\Sigma}_1 \xrightarrow{a.s.} \Sigma_1$ as $m \rightarrow \infty$.
Because

$$S^* = \frac{1}{n} \sum_{i=1}^n x_i x_i',$$

where

$$x_i \sim N_p(0, \Gamma + \sigma^2 \Sigma_1); i = 1, 2, \dots, n,$$

using Theorem 3.4.4 of Anderson (1984), p.81, the limiting distribution of

$$C(n) = \sqrt{n} (\hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}} - \Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p)$$

is normal with mean 0 and covariance

$$E(C_{ij}(n) C_{kl}(n)) = \sigma_{ik} \sigma_{jl} + \sigma_{il} \sigma_{jk} \quad (4.1)$$

where $\sigma_{ij} = (i, j)^{th}$ element of

$$\Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p.$$

(4.1) is obtained due to the fact that

$$S^{**} = \hat{\Sigma}_1^{-\frac{1}{2}} S^* \hat{\Sigma}_1^{-\frac{1}{2}} = \frac{1}{n} \sum_{i=1}^n u_i^* u_i^{*'} \quad \sim \quad \sim$$

asymptotically (using 3.1) and

$$u_i^* = \Sigma_1^{-\frac{1}{2}} x_i \sim N_p(0, \Sigma_1^{-\frac{1}{2}} \Gamma \Sigma_1^{-\frac{1}{2}} + \sigma^2 I_p) \quad \sim \quad \sim$$

From (2.13), estimator of $\Gamma + \sigma^2 \Sigma_1$ is

$$\begin{aligned}
 \hat{\Gamma} + \hat{\sigma}^2 \Sigma_1 &= \hat{\Sigma}_1^{-\frac{1}{2}} \tilde{B} (\Theta - \hat{\sigma}^2 I_q) \tilde{B}' \hat{\Sigma}_1^{-\frac{1}{2}} + \hat{\sigma}^2 \hat{\Sigma}_1 \\
 &= \hat{\Sigma}_1^{-\frac{1}{2}} \left(\sum_{i=1}^q (\alpha_i - \hat{\sigma}^2) w_i w_i' \right) \hat{\Sigma}_1^{-\frac{1}{2}} + \hat{\sigma}^2 \hat{\Sigma}_1
 \end{aligned}$$

where α_i 's, w_i 's and $\hat{\sigma}^2$ are explained in (2.13).

Because

$$\hat{\Sigma}_1 \xrightarrow{a.s.} \Sigma_1 \text{ as } m \rightarrow \infty \quad (\text{ using (3.1) })$$

$$w_i \xrightarrow{a.s.} d_i; i = 1, 2, \dots, q \text{ as } n \rightarrow \infty \quad (\text{ using (3.4) })$$

and

$$\hat{\sigma}^2 \xrightarrow{a.s.} \sigma^2 \text{ as } n \rightarrow \infty \quad (\text{ using (3.5) },$$

the limiting distribution of $\hat{\Gamma} + \hat{\sigma}^2 \Sigma_1$ is same as that of

$$\begin{aligned}
 & \Sigma_1^{-\frac{1}{2}} \left(\sum_{i=1}^q (\alpha_i - \sigma^2) d_i d_i' \right) \Sigma_1^{-\frac{1}{2}} + \sigma^2 \Sigma_1 \\
 & (\text{see Rao, 1983, p.122, (x)(b) }) \quad (4.2)
 \end{aligned}$$

Using the result of Anderson (1984) p.468,

$$E(\alpha_i) = l_i; i = 1, 2, \dots, q$$

asymptotically. Hence, from (4.2),

$$\begin{aligned}
 & E(\Sigma_1^{-\frac{1}{2}} (\sum_{i=1}^q \alpha_i - \sigma^2) \underset{\sim}{d_i} \underset{\sim}{d_i'} \Sigma_1^{\frac{1}{2}} + \sigma^2 \Sigma_1) \\
 &= \Sigma_1^{-\frac{1}{2}} (\sum_{i=1}^q (l_i - \sigma^2) \underset{\sim}{d_i} \underset{\sim}{d_i'}) \Sigma_1^{\frac{1}{2}} + \sigma^2 \Sigma_1 \\
 &= \Gamma + \sigma^2 \Sigma_1 \text{ (see 3.6 and 3.9).}
 \end{aligned}$$

From (4.2), the asymptotic variance of the estimator is same as that of

$$\sum_{i=1}^q \alpha_i \underset{\sim}{f_i} \underset{\sim}{f_i'}, \text{ where } \underset{\sim}{f_i} = \Sigma_1^{-\frac{1}{2}} \underset{\sim}{d_i} \quad (4.3)$$

From the result of Anderson (1984) p.468,

$\sqrt{n}(\alpha_i - l_i)$; $i = 1, 2, \dots, q$ are independently distributed and

$$\sqrt{n}(\alpha_i - l_i) \sim N(0, 2l_i^2) ; i = 1, 2, \dots, q \quad (4.4)$$

Hence, asymptotic variance of $\sum_{i=1}^q \alpha_i \underset{\sim}{f_i} \underset{\sim}{f_i'}$ can be obtained using (4.4). Call the asymptotic variance as

$$V(\sum_{i=1}^q \alpha_i \underset{\sim}{f_i} \underset{\sim}{f_i'}) = B. \quad (4.5)$$

References

- Anderson, T. W. (1963). Asymptotic theory for principal component analysis, *Annals of Mathematical Statistics*, 34, 122-138.
- Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis* (2nd Ed.). NY: Wiley.
- Bellman, R. (1960). *Introduction to matrix analysis*. New York: McGraw-Hill.
- Chen, P. (2002). A selection procedure for estimating the number of signal components. *Journal of Statistical Planning and Inference*, 105, 299-301.
- Chen, P., Wicks, M.C., & Adve, R. S. (2001). Development of a statistical procedure for detecting the number of signals in a radar measurement. *IEEE Proceedings of Radar, Sonar and Navigations*, 148(4), 219-226.
- Krishnaiah, P.R. (1976). Some recent developments on complex multivariate distributions, *Journal of Multivariate Analysis*, 6, 1-30.
- Kundu, D. (2000). Estimating the number of signals in the presence of white noise. *Journal of Statistical Planning and Inference*, 90, 57-68.
- Lawley, D. N., & Maxwell, A.E. (1963). *Factor Analysis as a Statistical Method*, Butterworths, London.
- Rao, C.R. (1983). Likelihood ratio tests for relationships between two covariance matrices. In: T. Amemiya, S. Karlin and L. Goodman, eds. *Studies in Econometrics, Time Series and Multivariate Statistics*. NY: Academic Press.
- Rao, C.R. (1983). *Linear statistical inference and its applications*. NY: Wiley Eastern Limited.
- Wax, M., T. Kailath (1985). Determination of the number of signals by information theoretic criteria, *IEEE Trans. Acoustics Speech Signal Processing ASSP-33*, 387-392.
- Wax, M., Shan, T. J., & Kailath, T. (1984). Spatio temporal spectral analysis by eigen structure methods, *IEEE Trans. Acoustics Speech Signal Processing ASSP-32*, 817-827.
- Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986a). On detection of number of signals in presence of white noise, *Journal of Multivariate Analysis*, 20, 1-25.
- Zhao, L. C., Krishnaiah, P. R., & Bai, Z. D. (1986b). On detection of the number of signals when the noise covariance matrix is arbitrary, *Journal of Multivariate Analysis*, 20, 26-49.

Utility of Weights for Weighted Kappa as a Measure of Interrater Agreement on Ordinal Scale

Moonseong Heo
Albert Einstein College of Medicine

Kappa statistics, unweighted or weighted, are widely used for assessing interrater agreement. The weights of the weighted kappa statistics in particular are defined in terms of absolute and squared distances in ratings between raters. It is proposed that those weights can be used for assessment of interrater agreements. A closed form expectations and variances of the agreement statistics referred to as AI_1 and AI_2 , functions of absolute and squared distances in ratings between two raters, respectively, are obtained. AI_1 and AI_2 are compared with the weighted and unweighted kappa statistics in terms of Type I Error rate, bias, and statistical power using Monte Carlo simulations. The AI_1 agreement statistic performs better than the other agreement statistics.

Key words: Kappa statistic, interrater agreement, bias, Type I Error rate, statistical power

Introduction

Kappa statistics, unweighted (Cohen, 1960) or weighted (Cohen, 1968), are used to measure interrater agreement. The unweighted kappa statistic is designed to measure agreement in nominal categorical ratings (Kraemer et al, 2002). Nevertheless, it is widely applied to agreement in ordinal ratings in medical research (e.g, Nelson & Pepe, 2000; Sim & Wright, 2005). In contrast, the weighted kappa statistics measure agreement in ordinal discrete ratings because it takes distances in ratings among raters into account (Fleiss et al., 2003).

The kappa statistics weighted and unweighted alike quantify observed agreement corrected for chance-expected agreement, and range from -1 to 1 . However, they are known to be sensitive to the marginal probabilities, e.g., prevalence in the diagnosis setting (Brennan &

Silman, 1992; Byrt et al., 1993). For instance, in a very special situation where all subjects have the characteristic that is being assessed, the kappa statistics may not necessarily be informative. Suppose that a rating scale or instrument item measures a psychotic feature of subjects with ratings 0 for absence and 1 for the presence of the feature. If the instrument has a perfect sensitivity, all of well-trained raters would rate 1 for the subjects when all the subjects have that particular psychotic feature. In this situation, the kappa statistics are undefined based on its formula because both the numerator and the denominator are 0 .

With respect to the sign of the kappa statistics, it does not necessarily serve as an indicator for direction of agreement. For instance, a negative kappa does not necessarily indicate that raters disagree in ratings. But it only indicates by definition that chance-expected agreement is greater than observed agreement. On the other hand, the kappa statistics can return a positive agreement even when observed disagreement overwhelms by far observed agreement, implying again by definition that a positive kappa does not necessarily mean that raters agree in ratings. Thus, the kappa statistics return a positive value no matter how small the observed agreement is as long as it exceeds agreement expected by chance. At the same

Moonseong Heo is Associate Professor in the Department of Epidemiology and Population Health at the Einstein College of Medicine. He is interested in longitudinal data analysis and sample size estimations in designing clinical trials with repeated measures. Email: moonseong.heo@einstein.yu.edu

time, it is possible to have a low kappa for high agreements as discussed in Feinstein and Cicchetti (1990a, 1990b). For these reasons, some argued that the kappa statistics are a measure of association rather than that of agreement (Graham and Jackson, 1993).

In this article, we explored the utility of the weights that have been used for the weighted kappa statistics as alternative agreement statistics (rather than as a measure of association) to complement such undesirable features of the kappa statistics in certain, if not general, situations. Vast amount of literature has been devoted to discussion of kappa statistics (for reviews e.g., Maclure & Willett, 1987; Agresti 1992; Kraemer, 1992; Shrout, 1998; Banerjee et al, 1999) and other types of alternative agreement measures have been proposed (e.g., O'Connell & Dobson, 1984; Kuper and Hafner, 1989; Aickin, 1990; Uebersax, 1993; Donner & Eliasziw, 1997). Nevertheless, the utility of the weights has not been discussed in the literature.

Two agreement statistics are investigated, which are averages of observed weights defined in terms of distances in ratings between two raters and quantify a degree of agreement compared to the possibly worst disagreement. Sampling distributions of those two agreement statistics are derived and compared with those of the unweighted and weighted kappa statistics with respect to Type I Error rate, bias of sample estimates and variances, and statistical power under various scenarios. Monte Carlo simulations were used to conduct the comparisons.

Methods

Agreement Statistics

Assume that two raters rate N subjects using an instrument with K ordinal ratings denoting the i -th rater's rating for the j -th subject R_{ij} ; $i = 1, 2$; $j = 1, \dots, N$; the ordinal rating R ranges from 1 to K by 1.

Unweighted kappa statistic

The (unweighted) kappa is a function of observed and chance-expected agreements in categorical ratings between raters. As described in Fleiss et al (2003), the observed agreement

can be quantified by

$$p_o = \sum_{k=1}^K P(R_{1i} = R_{2j} = k) = \sum_{j=1}^N 1(R_{1j} = R_{2j}) / N \quad (1)$$

and the chance-expected agreement by

$$p_e = \sum_{k=1}^K p_{1k} p_{2k} \quad (2)$$

where $1(x)$ is an indicator function which returns 1 if the condition x is met and 0 otherwise, and

$$p_{ik} = P(R_i = k) = \sum_{j=1}^N 1(R_{ij} = k) / N \quad (3)$$

is the marginal probability of the i -th rater's rating being k . The kappa statistic κ is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (4)$$

This formula indicates that the kappa statistic represents the difference in probability between the observed (1) and chance-expected (2) agreement (the numerator) relative to the complement of the expected agreement (the denominator). Although the kappa statistic (4) ranges from -1 to 1 , its sign does not necessarily indicate a direction of agreement.

Weighted kappa statistics

Weighted kappa has also been proposed to reflect relative seriousness of disagreement between raters (Cicchetti, 1976). Interrater disagreement can be quantified as absolute or squared distance in ordinal ratings. Thus, two typical weights that are used for calculating weighted kappa statistics are as follows:

$$w_{kk'}^{(1)} = 1 - \frac{|k - k'|}{(K - 1)} \quad (5)$$

(Cicchetti & Allison, 1971) and

$$w_{kk'}^{(2)} = 1 - \frac{(k - k')^2}{(K - 1)^2} \quad (6)$$

(Fleiss & Cohen, 1973) where k and k' are rater's ratings such that $R_1 = k$ and $R_2 = k'$. It is obvious that: 1) both weights range from 0 to 1 because the denominator $(K - 1)$ or $(K - 1)^2$ represent the worst disagreement; 2) the ratings should be ordinal in order for the weights to represent meaningful disagreements (distances in nominal ratings have little meaning with respect to disagreement.) Subsequently, weighted kappa statistics can be obtained in a similar manner to the unweighted kappa (4) as follows:

$$\kappa_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}} \quad (7)$$

where

$$p_{o(w)} = \sum_{k=1}^K \sum_{k'=1}^K w_{kk'} P(R_1 = k, R_2 = k')$$

and

$$p_{e(w)} = \sum_{k=1}^K \sum_{k'=1}^K w_{kk'} p_{1k} p_{2k'};$$

and

$$P(R_1 = k, R_2 = k') = \sum_{j=1}^N 1(R_{1j} = k, R_{2j} = k') / N.$$

Denote $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$ for the weighted kappa statistics when $w = w^{(1)}$ and $w^{(2)}$, respectively. The weighted kappa (7) also ranges from -1 to 1 , representing only the difference in observed and chance expected agreement without bearing of direction. Of note, the weighted kappa $\kappa_{w^{(2)}}$ is the same as the intraclass correlation coefficient (ICC, Bartko, 1966; Shrout & Fleiss, 1979) aside from a term involving the factor $1/N$ (Fleiss and Cohen, 1973). Further, the unweighted kappa statistic (4) is a special case of a weighted kappa when $w_{kk'} = 1 (k = k')$. Especially when $K = 2$, both $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$ are the same as the unweighted kappa statistic (4).

Agreement Index, AI, based on the weights

The weights $w^{(1)}$ (5) and $w^{(2)}$ (6) *per se* can be used for measurement of interrater agreement because the weights represent degrees of (dis)agreement in rating distances between raters on each individual subject in a normalizing manner—*normalization by the possibly worst disagreements*. Therefore, it is proposed that the averages of observed weights *over the subjects* can serve as alternative agreement statistics. Denote them by AI_1 and AI_2 for “Agreement Index” as follows:

$$AI_1 = 1 - \frac{\sum_{j=1}^N |R_{1j} - R_{2j}|}{N(K - 1)} \quad (8)$$

and

$$AI_2 = 1 - \frac{\sum_{j=1}^N (R_{1j} - R_{2j})^2}{N(K - 1)^2} \quad (9)$$

It is apparent that both agreement indices AI_1 and AI_2 range from 0 to 1. It will be shown in the next section that: the closer the indices are to 0, the stronger the degree of disagreement; the closer to 1, the greater the extent of agreement. When $K = 2$, AI_1 and AI_2 are identical to each other because the absolute and squared distances are the same between 0 and 1, and are the same as the observed agreement p_o in equation (1).

Sampling Distributions

The AI Statistics

The sampling distributions of the AI statistics are presented under a null situation where the following two conditions are met:

- Condition A. (“Marginal equal probability” condition): Ratings are *marginally* uniform in multinomial probability, i.e., $P(R_{ij}=k) = 1/K$, for all i, j , and k ;
- Condition B. (“Joint independent rating” condition): The two rater ratings R_1 and R_2 are *jointly* independent, i.e., $P(R_1=k, R_2=k') = P(R_1=k)P(R_2=k')$.

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Condition A reflects a situation that both raters assess the subject in a uniform and blinded manner. In that the marginal probability distribution of the subjects' true ratings does not depend on the raters (as it should not by definition) unlike that of the kappa statistics, which relies on the rater-dependent estimates of marginal probabilities as reflected in equation (3). Condition B reflects a situation where the two raters assess independently as is the case for the kappa statistics.

When taken together, therefore, the combination of both condition A and B represents a null situation where the observed agreement between raters is purely random with no opportunity for any systematic agreement. Departure from either condition will be an alternative non-null situation of systematic agreement or disagreement.

Under the null situation with both conditions A and B, the first two sampling moments of AI_1 and AI_2 can be derived based on the following probability of distances in ratings between the two raters:

$$\begin{aligned} P(|R_1 - R_2| = d) &= \sum_{|r_1 - r_2| = d} P(R_1 = r_1, R_2 = r_2) = \\ &= \sum_{|r_1 - r_2| = d} P(R_1 = r_1) P(R_2 = r_2) = \\ &= 2(K - d) / K^2 \end{aligned}$$

It follows that:

$$\begin{aligned} E|R_1 - R_2|^n &= \sum_{d=1}^{K-1} d^n P(|R_1 - R_2| = d) = \\ &= \frac{2}{K^2} \sum_{d=1}^{K-1} (K - d) d^n. \end{aligned}$$

Hence

$$\begin{aligned} E|R_1 - R_2| &= \frac{K^2 - 1}{3K}, \\ Var|R_1 - R_2| &= E|R_1 - R_2|^2 - \\ &= \frac{(K^2 + 2)(K^2 - 1)}{18K^2} \end{aligned}$$

and

$$\begin{aligned} E(R_1 - R_2)^2 &= \frac{K^2 - 1}{6} \\ Var(R_1 - R_2)^2 &= E|R_1 - R_2|^4 - \\ &= \frac{7K^4 - 20K^2 + 13}{180} \end{aligned}$$

Thus, under the null situation: for AI_1 ,

$$E(AI_1) = \frac{2K - 1}{3K}, \quad (10)$$

$$Var(AI_1) = \frac{(K + 1)(K^2 + 2)}{18NK^2(K - 1)}; \quad (11)$$

and for AI_2 ,

$$E(AI_2) = \frac{5K - 7}{6(K - 1)}, \quad (12)$$

$$Var(AI_2) = \frac{7K^4 - 20K^2 + 13}{180N(K - 1)^4}. \quad (13)$$

The expected $E(AI_1)$ and $E(AI_2)$ (Table 1) represent chance expected agreement similar to the notion of p_e (2) of the kappa statistic. Thus, observed AI 's less than expected $E(AI)$'s indicate systematic (as opposed to purely random) disagreement between raters because observed distances in disagreement is larger than what is expected under the conditions A and B. Subsequently, normal-approximated test statistics

$$Z_{AI_1} = (AI_1 - E(AI_1)) / se(AI_1) \quad (14)$$

and

$$Z_{AI_2} = (AI_2 - E(AI_2)) / se(AI_2) \quad (15)$$

can be used for testing significance of interrater agreement and for direction of systematic agreement as well.

The kappa statistics

Derivation of sampling distribution of the un- and weighted kappa statistics under a null situation is based only on condition B.

These kappa statistics, (4) and (7), use the rater-dependent marginal probability distributions of the subjects for derivation of their samplings distributions. The expected kappa statistic under condition B is 0. The standard error (se) of kappa is under condition B known as:

$$\text{se}(\kappa) = \frac{1}{\sqrt{N(1-p_e)}} \sqrt{p_e + p_e^2 - \sum_{k=1}^K p_{1k} p_{2k} (p_{1k} + p_{2k})} \quad (16)$$

(Fleiss et al., 1969). From this, a normal-approximated test statistic

$$z_{\kappa} = \kappa / \text{se}(\kappa) \quad (17)$$

is used to test significance of agreement between two raters, i.e. $H_0: \kappa = 0$.

The expected weighted kappa statistic under condition B is also 0. The standard error (se) of weighted kappa under condition B has the following formula as described elsewhere (Fleiss et al., 1969; Cicchetti & Fleiss, 1977; Landis & Koch, 1977; Fleiss & Cicchetti, 1978; Huber, 1978):

$$\text{se}(\kappa_w) = \frac{1}{\sqrt{N(1-p_{e(w)})}} \sqrt{\sum_{k=1}^K \sum_{k'=1}^K p_{1k} p_{2k'} [w_{kk'} - (\bar{w}_{1k} + \bar{w}_{2k'})]^2 - p_{e(w)}^2} \quad (18)$$

where $\bar{w}_{1k} = \sum_{k'=1}^K p_{2k'} w_{kk'}$ and $\bar{w}_{2k'} = \sum_{k=1}^K p_{1k} w_{kk'}$.

Both normal-approximated test statistics,

$$z_{\kappa_w^{(1)}} = \kappa_{w^{(1)}} / \text{se}(\kappa_{w^{(1)}}) \quad (19)$$

and

$$z_{\kappa_w^{(2)}} = \kappa_{w^{(2)}} / \text{se}(\kappa_{w^{(2)}}), \quad (20)$$

are used for testing significance of interrater agreement, that is testing $H_0: \kappa_w = 0$.

Simulation Design and Evaluation Measures for Comparisons

Simulation Design

For evaluations under null situations, the parameters considered are $K = 2, 3, 4, 5$ and $N = 20, 30, 40, 50, 100, 200$. For each combination of K and N , generated 10,000 simulated datasets of ratings from two raters from multinomial distributions meeting both conditions A and B, i.e., the joint probabilities of the ratings are $P(R_1=k, R_2=k') = P(R_1=k)P(R_2=k') = 1/K^2$ for all k, k' , and K .

For evaluations under alternative (referred to as a departure from null) situations, consider 6 alternative situations where both conditions A and B are not met when $K = 3$. The joint probabilities of ratings between two raters are represented in 6 configurations in Table 2. From a joint multinomial distribution with those $K^2 = 9$ probabilities specified for each configuration, randomly generated ratings between two raters. Configuration 4 in particular represents a situation where condition B is met but condition A is not. For each configuration, we considered $N = 20, 30, 40$, and 50 , and generated 10,000 datasets.

The simulations were conducted using S-plus v6.2 statistical software. In empirical comparisons of the five agreement statistics (κ , $\kappa_{w^{(1)}}$, $\kappa_{w^{(2)}}$, AI_1 and AI_2), the following evaluation measures were used: percent bias in sample estimates and variances, Type I Error rate, and statistical power.

Evaluation measures for bias in sample estimates

The percent biases in sample estimates of the two AI statistic, (8) and (9), are obtained as follows:

$$\% \text{Bias in sample estimates} = \frac{\overline{AI} - E(AI)}{E(AI)} \times 100,$$

where $\overline{AI}$ is the sample estimate of an AI statistics, i.e., $\overline{AI} = \sum_{s=1}^{N_{sim}} AI(s) / N_{sim}$; $AI(s)$ represents the s -th estimate of an AI from $N_{sim} =$

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

10,000 simulations; $E(AI)$ is defined in equations (10) and (12). The corresponding percent biases in sample estimates of the kappa statistics were undefined because expectations under the null are all zero.

Evaluation measures for bias in sample variances

The percent biases in sample variances are computed as follows. First, for the AI statistics,

$$\begin{aligned} &\% \text{Bias in sample variance of } AI \\ &= \frac{\hat{Var}(AI) - Var(AI)}{Var(AI)} \times 100, \end{aligned}$$

where

$$\hat{Var}(AI) = \left[\sum_{s=1}^{N_{sim}} AI^2(s) - N_{sim} \overline{AI}^2 \right] / N_{sim}$$

is the sample estimates of variance of an AI statistics from 10,000 simulations, and $Var(AI)$ is defined in equations (11) and (13).

Second, for the three kappa statistics, (4) and (7),

$$\begin{aligned} &\% \text{Bias in sample variance of kappa} = \\ &= \frac{\hat{Var}(\kappa) - \overline{Var(\kappa)}}{Var(\kappa)} \times 100, \end{aligned}$$

where the term

$$\hat{Var}(\kappa) = \left[\sum_{s=1}^{N_{sim}} \kappa^2(s) - N_{sim} \bar{\kappa}^2 \right] / N_{sim}$$

in numerator represents the sample estimate of variance of a kappa statistic; and

$\overline{Var(\kappa)} = \sum_{s=1}^{N_{sim}} Var_s(\kappa) / N_{sim}$ is the sample

average of a variance $Var(\kappa)$, the square of a standard error, (16) or (18), of a kappa statistic;

and $\bar{\kappa} = \sum_{s=1}^{N_{sim}} \kappa_s / N_{sim}$ is the sample estimate of

a kappa statistic. All of these are obtained from $N_{sim}=10,000$ simulations.

Evaluation measures for type I error rated and power

Type I Error rates and statistical power were obtained as proportions of p-values (obtained from the standard normal z tests, (14), (15), (17), (19) and (20)) less than a 0.05 nominal significance level from 10,000 simulations under the null and alternative situations, respectively, as described above.

Results

Null situations

Bias in sample mean: Table 3(a) shows averages of the agreement statistics over 10,000 simulations and their %bias (The %biases of (un)weighted kappa statistics, (4) and (7), were not computed because their expected values are zero under the null situation.) As can be seen, the %bias is minimal for all the agreement statistics; all of absolute %bias is less than 0.4%.

Bias in sample variance: Table 3(b) shows that %bias in estimated variances of the agreement statistics are also very small. However, % biases of variances of the kappa statistics (absolute %bias <8.1%) are larger than those of the AI statistics (absolute %bias <3.2%).

Type I Error rate: Table 3(c) shows that type I error rates of the five agreement statistics are fairly close to the nominal alpha-level 0.05 over the combinations of K and N considered here.

Alternative situations

Configuration 1 (Symmetric agreement):

This configuration represents an ideal pattern of agreements between two raters. Two raters agree equally on each rating and disagreement reduces, as the differences in ratings get larger. All the five agreements show positive agreements (Table 4(a)) and high statistical powers even when N is as small as 30 (Table (b)) with the 60% observed agreement. Overall, AI_1 showed the greatest power.

Configuration 2 (Triangular):

This configuration represents a situation where one rater's ratings are always no less than those of the other. Further, a rather extreme situation was considered where the observed agreement is as small as 15%. All of the kappa statistics returns positive value, albeit small, implying that the

observed agreement is beyond the chance expected agreement (Table 4 (a)).

Conversely, the other two AI statistics returned value much smaller than expected under null, implying that the two raters systematically disagree. The statistical power of the unweighted kappa is relatively much higher (about 40% for $N=50$) compared to that of the other weighted kappa (less than 11% for the same N ; Table 4(b)). The statistical power of the AI statistics are near perfect even with $N=20$ implying strong disagreement between the two raters. Overall AI_2 showed the greatest power, slightly larger than that of AI_1 .

Configuration 3 (Skewed): This configuration represents where major agreement occurs at one rating; in this case, the rating is 3. The observed agreement is 72% where 68% observed agreements accounts for $R=3$ and the other 4% for $R=1$ and 2. All of the five agreement statistics showed positive agreement (Table 4(a)). However, the statistical power of the three kappa statistics is much smaller (at about 40% for $N=50$) than that of AI_1 (over 85% for $N=20$). The statistical power of AI_2 was in between them but toward AI_1 for larger N .

Configuration 4 (Independent): This configuration represents a situation where ratings between raters are independent but not in a uniform manner with 54% observed agreement. In other words, this configuration satisfies the null condition B but not A as mentioned before. Table 4(a) shows that the three kappa statistics are all near around 0 as expected. However, the AI statistics were greater than what is expected under the null situation. With respect to statistical power, the kappa statistics returned power around the nominal level 0.05 as also expected. On the other hand, both AI statistics returned greater power. Overall, AI_1 showed the greatest power.

Configuration 5 (Incomplete): This configuration represents a situation where both raters rated only 2 and 3 with 75% observed power. This often happens not because the raters are biased or informed a priori but because the study subjects were recruited based on particular exclusion/inclusion criteria, which may rule out category 1 of an instrument item. In this case, the kappa statistics behave the same way with only two ratings available, i.e., $K=2$. This is

reflected on Table 4(a) and (b) in that the three kappa statistics have the same kappa values as well as the same statistical power. However, their power is much smaller than that of both AI 's (Table 4(b)), perhaps because these AI 's are based on $K=3$ rather than $K=2$.

Configuration 6 (Symmetric disagreement): This configuration represent a "systematic" disagreement between two raters in that the off-diagonal disagreement proportion gets larger away from the diagonal agreement. The observed agreement in this configuration is 15%, which is the same as that of configuration 2, in which the kappa statistics were positive. Under the present configuration, all the three kappa statistics returned negative values still not necessarily implying in theory that the raters disagree. Both AI statistics are smaller than what is expected under the null, implying that the raters systematically disagree. The statistical power of the kappa statistics is comparable with that of AI 's for larger N . Overall, however, AI_1 showed the greatest power.

Bias of variance of the agreement statistics: Table 4(c) shows %bias of the variance estimates of the five agreement statistics. The negative %bias indicates that variance estimate under alternative situations are smaller than that under the null situation. Because the square root of variance under the null was used for the denominators of the z -test statistics ((14), (15), (17), (19) and (20)), tests with negative %bias of variance estimates under alternative situations are conservative. It follows that the z -test of AI_1 is the most conservative test. Despite this, AI_1 returned the greatest power under almost configurations (Table 4(b)).

Discussion

The overall finding from this study is that AI_1 and AI_2 statistics, (8) and (9), based on the weights that have been used for calculation of weighted kappa are useful agreement statistics. Specifically, compared with the other agreement statistics, AI_1 in particular has desirable properties in terms of type I error, bias in mean and variance, sensitivity in direction of agreement and statistical power.

The expectation and variance of AI_1 and AI_2 under the null situation have closed form

expression $E(AI)$ in equations (10) and (12), and $Var(AI)$ in equations (11) and (13), and thus are ready to be used for sample size calculation for pre-specified power and K , the number of ratings. Both AI_1 and AI_2 are capable for any kind of combination of rater ratings, even when two raters rated only one particular rating across all subjects, a “single cell” situation. In this case, any kappa statistic is not defined and at the same time ICC is also uninformative because of no variation of rating over the subjects, i.e., zero total variation. In the single cell situation, both AI_1 and AI_2 will always be 1 as long as the single cell falls onto a diagonal cell. If it falls onto farthest northeast or southwest corner, then both will be 0. Otherwise, they will depend on K .

The weighted kappa statistics, $\kappa_{w^{(1)}}$ and $\kappa_{w^{(2)}}$, did not appear to have sizable advantage over the unweighted kappa statistic. This is somewhat surprising because the weights *per se*, AI_1 (in particular) and AI_2 , perform much better than the unweighted kappa statistic. This may be due to a discrepancy in viewpoints on agreement between the kappa statistics and the AI statistics. In short, the kappa statistics are based on probabilities particularly focusing on whether or not the inter rater ratings are “independent.”

In contrast, the AI statistics are based on distances in ratings between two raters regardless of independence. The normalization of the distances against the possibly worst distance implies that the AI statistics are indeed goodness-of-fit indices, a different view from that of the kappa statistics. Another discrepancy is also reflected on the null situations. Indeed, the null situation (both conditions A and B) of the AI statistics is a special case of the null situation (only condition B) of the kappa statistics. It is an open question and debatable which null situation should be adopted in agreement assessment.

Both AI_1 and AI_2 can easily be extended to cases for multiple raters ($i=1, \dots, I$) as follows:

$$AI_1^m = 1 - \frac{\sum_{i=1}^I \sum_{i'=i+1}^I \sum_{j=1}^N |R_{ij} - R_{i'j}|}{I(I-1)N(K-1)/2}$$

and

$$AI_2^m = 1 - \frac{\sum_{i=1}^I \sum_{i'=i+1}^I \sum_{j=1}^N (R_{ij} - R_{i'j})^2}{I(I-1)N(K-1)^2/2}$$

Note that AI_1 and AI_2 are special cases of AI_1^m and AI_2^m , respectively, for $I = 2$. Expectations of AI_1^m and AI_2^m are the same as those of AI_1 and AI_2 , respectively. However, derivation of variances of AI_1^m and AI_2^m are cumbersome because they are not a sum of independent distances. Nevertheless, the variances can empirically be derived by use of Monte Carlo simulations under the null situation. These empirically obtained variances can consequently be used for testing significance of agreement among multiple raters. Furthermore, in computation of AI_1^m and AI_2^m , it is not required that all raters rate every subject. In the presence of missing ratings, the denominators AI_1^m and AI_2^m will be adjusted to the number of available distances.

Although not explored in the present article, Lipsitz et al. (1994) considered a marginal and a joint probability distribution of two ratings (positive vs. negative) to derive a class of estimators for kappa using an estimating equation. In that they compared their estimating equation estimators to maximum likelihood estimator (MLE) obtained under a beta-binomial distribution derived by Verducci et al (1988). However, validity of both estimating equation estimators and MLE relies on a large sample size (Fleiss et al, 2003). Small sample properties were discussed in Koval and Blackman (1996) and Gross (1986).

In conclusion, both AI_1 and AI_2 are sensitive to the magnitude as well as the direction of agreement between two raters, and generally have greater power relative to the kappa statistics. Thus, both AI_1 and AI_2 can serve as agreements statistics of their own as well as complement statistics to the kappa statistics.

References

Agresti A (1992) Modelling patterns of agreement and disagreement. *Statistical Methods in Medical Research* 1, 201-218.

Aickin M (1990) Maximum likelihood estimation of agreement in the constant predictive probability model, and its relation to Cohen's kappa. *Biometrics* 46, 293-302.

Bartko JJ (1966) The intraclass correlation coefficient as a measure of reliability. *Psychological Reports* 19, 3-11.

Banerjee M, Capozzoli M, McSweeney L (1999) Beyond kappa: A review of interrater agreement measure. *Canadian Journal of Statistics* 27, 3-23.

Brennan RI, Silman A (1992) Statistical Methods for assessing observer variability in clinical measures. *British Medical Journal* 304, 1491-1494.

Byrt T, Bishop J, Carlin JB (1993) Bias, prevalence and kappa. *Journal of Clinical Epidemiology* 46, 423-429.

Cicchetti DV, Allison T (1971) A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* 11, 101-109.

Cicchetti DV (1976) Assessing interrater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry* 129, 452-456.

Cicchetti DV, Fleiss JL (1977) Comparison of the null distribution of weighted kappa and the C ordinal statistics. *Applied Psychological Measurement* 1, 195-201.

Cohen J (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1, 195-201.

Cohen J (1968) Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* 70, 213-220.

Donner A, Eliasziw M (1997) A hierarchical approach to inferences concerning interobserver agreement for multinomial data. *Statistics in Medicine* 16, 1097-1106.

Feinstein AR, Cicchetti DV (1990a) High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 543-549.

Feinstein AR, Cicchetti DV (1990b) High agreement but low kappa: II. The problems of two paradoxes. *Journal of Clinical Epidemiology* 43, 551-558.

Fleiss JL, Cohen J, Everitt BS (1969) Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin* 72, 323-327.

Fleiss JL, Cohen J (1973) The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement* 33, 613-619.

Fleiss JL, Cicchetti DV (1978) Inference about weighted kappa in the non-null case. *Applied Psychological Measurement* 2, 113-117.

Fleiss JL, Levin B, Paik MC (2003) *Statistical Methods for Rates and Proportions*, 3rd ed., New York: Wiley, Ch. 18.

Graham P, Jackson R (1993) The analysis of ordinal agreement data: Beyond weighted kappa. *Journal of Clinical Epidemiology* 46, 1055-1062.

Gross ST (1986). The kappa coefficient of agreement for multiple observers when the number of subjects is small. *Biometrics* 42, 883-893.

Hubert LJ (1978) A general formula for the variance of Cohen's weighted kappa. *Psychological Bulletin* 85, 183-184.

Koval JJ, Blackman NJM (1996) Estimators of kappa-exact small sample properties. *Journal of Statistical Computations and Simulations* 55, 315-336.

Kraemer HC (1992) Measurement of reliability for categorical data in medical research. *Statistical Methods in Medical Research* 1, 183-199.

Kraemer HC, Periyakoil VS, Noda A (2002) Tutorial in Biostatistics: Kappa coefficients in medical research. *Statistics in Medicine* 21, 2109-2129.

Kupper LL, Hafner KB (1989) On assessing interrater agreement for multiple attribute responses. *Biometrics* 45, 957-967.

Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33, 159-174.

Lipsitz SR, Laird NM, Brennan TA (1994) Simple moment estimates of the κ -coefficient and its variance. *Applied Statistics* 43, 309-323.

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Maclure M, Willett WC (1987). Misinterpretation and misuse of the kappa statistic. *American Journal of Epidemiology* 126, 161-169.

Nelson JC, Pepe MS (2000) Statistical description of interrater variability in ordinal ratings. *Statistical Methods in Medical Research* 9, 475-496.

O'Connell DL, Dobson AJ (1984) General observer-agreement measures on individual subjects and groups of subjects. *Biometrics* 40, 973-983.

Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin* 86, 420-428.

Shrout PE (1998) Measurement reliability and agreement in psychiatry. *Statistics Methods in Medical Research* 7, 301-317.

Sim J, Wright CC (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy* 85, 257-268.

Uebersax JS (1993) Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of American Statistical Association* 88, 421-427.

Verducci JS, Mack ME, DeGroot MH (1988). Estimating multiple rater agreement for a rare diagnosis. *Journal of Multivariate Analysis* 27, 512-535.

Appendix

Table 1: Expected $E(AI_1)$, $E(AI_2)$, $Var(AI_1)$, and $Var(AI_2)$

N	Quantity	K			
		2	3	4	5
	$E(AI_1)$	0.500	0.556	0.583	0.600
	$E(AI_2)$	0.500	0.667	0.722	0.750
20	$Var(AI_1) \times 1,000$	12.50	6.79	5.21	4.50
	$Var(AI_2) \times 1,000$	12.50	6.94	5.09	4.22
30	$Var(AI_1) \times 1,000$	8.33	4.53	3.47	3.00
	$Var(AI_2) \times 1,000$	8.33	4.63	3.40	2.81
40	$Var(AI_1) \times 1,000$	6.25	3.40	2.60	2.25
	$Var(AI_2) \times 1,000$	6.25	3.47	2.55	2.11
50	$Var(AI_1) \times 1,000$	5.00	2.72	2.08	1.80
	$Var(AI_2) \times 1,000$	5.00	2.78	2.04	1.69
100	$Var(AI_1) \times 1,000$	2.50	1.36	1.04	0.90
	$Var(AI_2) \times 1,000$	2.50	1.39	1.02	0.84
200	$Var(AI_1) \times 1,000$	1.25	0.68	0.52	0.45
	$Var(AI_2) \times 1,000$	1.25	0.69	0.51	0.42

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Table 2: Configurations of probability $P(R_1 = k, R_2 = k')$ used for the alternative situations under which the comparison of agreement measures was made: $K = 3$.

Configuration 1: Symmetric Agreement				Configuration 4: Independent			
R_1	R_2			R_1	R_2		
	1	2	3		1	2	3
1	0.20	0.08	0.04	1	0.01	0.02	0.07
2	0.08	0.20	0.08	2	0.02	0.04	0.14
3	0.04	0.08	0.20	3	0.07	0.14	0.49

Configuration 2: Triangular				Configuration 5: Incomplete			
R_1	R_2			R_1	R_2		
	1	2	3		1	2	3
1	0.05	0.10	0.65	1	0.000	0.000	0.000
2	0.00	0.05	0.10	2	0.000	0.150	0.125
3	0.00	0.00	0.05	3	0.000	0.125	0.600

Configuration 3: Skewed				Configuration 6: Symmetric disagreement			
R_1	R_2			R_1	R_2		
	1	2	3		1	2	3
1	0.02	0.02	0.06	1	0.050	0.100	0.225
2	0.02	0.02	0.06	2	0.100	0.050	0.100
3	0.06	0.06	0.68	3	0.225	0.100	0.050

Table 3(a): Comparison of agreement measures under the null situation: Sample Mean and Percent Bias from 10,000 simulations.

K	N	κ	$\mathcal{K}_{w^{(1)}}$	$\mathcal{K}_{w^{(2)}}$	AI_1	%bias	AI_2	%bias
2	20	-0.001	-0.001	-0.001	0.499	-0.1	0.499	-0.1
	30	0.002	0.002	0.002	0.501	0.2	0.501	0.2
	40	0.002	0.002	0.002	0.501	0.3	0.501	0.3
	50	0.003	0.003	0.003	0.501	0.3	0.501	0.3
	100	0.001	0.001	0.001	0.500	0.1	0.500	0.1
	200	0.000	0.000	0.000	0.500	0.0	0.500	0.0
3	20	0.000	-0.001	-0.003	0.555	-0.1	0.666	-0.2
	30	0.001	0.002	0.002	0.556	0.1	0.667	0.1
	40	0.003	0.003	0.003	0.557	0.2	0.668	0.2
	50	-0.001	-0.001	-0.001	0.555	-0.1	0.666	-0.1
	100	0.000	0.000	0.000	0.556	0.0	0.667	0.0
	200	0.000	0.000	-0.001	0.555	-0.1	0.666	-0.1
4	20	0.000	-0.002	-0.004	0.583	0.0	0.721	-0.1
	30	0.000	-0.001	-0.001	0.583	0.0	0.722	0.0
	40	-0.001	0.000	0.000	0.583	0.0	0.722	0.0
	50	-0.001	0.000	0.000	0.583	0.0	0.722	0.0
	100	0.000	-0.001	-0.001	0.583	0.0	0.722	0.0
	200	0.000	0.000	0.000	0.583	0.0	0.722	0.0
5	20	-0.001	-0.001	-0.001	0.599	0.0	0.750	-0.1
	30	0.001	0.002	0.002	0.601	0.0	0.750	0.1
	40	0.001	0.001	0.001	0.600	0.0	0.750	0.0
	50	-0.001	0.000	0.002	0.600	0.0	0.751	0.1
	100	-0.001	-0.001	-0.001	0.600	0.0	0.750	0.0
	200	0.000	0.000	0.000	0.600	0.0	0.750	0.0
Mean*		0.000	0.000	0.000		0.0		0.0
Median*		0.000	0.000	0.000		0.0		0.0
SD*		0.001	0.001	0.002		0.1		0.1

*Column Mean, Median, and SD.

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Table 3(b): Comparison of agreement measures under the null situation: Sample Variance and Bias from 10,000 simulations.

K	N	<i>Var</i>		<i>Var</i>		<i>Var</i>		<i>Var</i>		<i>Var</i>	
		<i>Var</i> (κ)	%bias	($\kappa_{w^{(1)}}$)	%bias	($\kappa_{w^{(2)}}$)	%bias	(AI_1)	%bias	(AI_2)	%bias
2	20	0.048	5.2	0.048	5.2	0.048	5.2	0.012	-0.2	0.012	-0.2
	30	0.033	4.5	0.033	4.5	0.033	4.5	0.008	1.4	0.008	1.4
	40	0.024	2.5	0.024	2.5	0.024	2.5	0.006	0.0	0.006	0.0
	50	0.019	0.7	0.019	0.7	0.019	0.7	0.005	-1.5	0.005	-1.5
	100	0.010	0.1	0.010	0.1	0.010	0.1	0.002	-0.9	0.002	-0.9
	200	0.005	-0.6	0.005	-0.6	0.005	-0.6	0.001	-1.1	0.001	-1.1
3	20	0.025	8.0	0.030	6.6	0.047	5.5	0.007	0.9	0.007	-0.3
	30	0.016	3.1	0.020	2.2	0.031	1.7	0.004	-1.7	0.005	-2.4
	40	0.012	1.3	0.015	0.2	0.024	0.1	0.003	-2.2	0.003	-1.7
	50	0.010	1.6	0.012	1.9	0.019	1.9	0.003	0.9	0.003	1.0
	100	0.005	0.5	0.006	1.5	0.010	1.7	0.001	1.0	0.001	0.9
	200	0.002	0.3	0.003	0.5	0.005	0.9	0.001	-1.5	0.001	-1.3
4	20	0.016	3.4	0.024	2.6	0.046	3.6	0.005	-1.5	0.005	-0.4
	30	0.011	4.3	0.017	2.7	0.032	2.3	0.003	-1.0	0.003	-1.3
	40	0.008	2.0	0.013	3.1	0.024	3.5	0.003	0.7	0.003	0.5
	50	0.007	3.9	0.010	2.7	0.020	2.2	0.002	1.7	0.002	1.8
	100	0.003	2.1	0.005	2.8	0.010	2.7	0.001	1.4	0.001	1.2
	200	0.002	-0.3	0.003	0.7	0.005	1.7	0.001	0.8	0.001	1.6
5	20	0.012	7.2	0.023	6.2	0.047	5.5	0.004	-0.2	0.004	-0.7
	30	0.008	3.9	0.015	4.3	0.032	4.3	0.003	-0.1	0.003	-0.8
	40	0.006	2.6	0.012	3.9	0.024	3.2	0.002	-0.3	0.002	-1.4
	50	0.005	1.3	0.009	1.7	0.019	1.9	0.002	-2.7	0.002	-3.1
	100	0.002	-0.2	0.005	-1.4	0.010	-2.0	0.001	-0.6	0.001	-0.6
	200	0.001	0.2	0.002	0.4	0.005	0.4	0.000	1.1	0.000	1.7
Mean*			2.4		2.3		2.2		-0.2		-0.3
Median*			2.1		2.4		2.1		-0.2		-0.5
SD*			2.3		2.1		2.0		1.2		1.3

*Column Mean, Median, and SD.

Table 3(c): Comparison of agreement measures under the null situation: Type I error rate from 10,000 simulations.

K	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
2	20	0.050	0.050	0.050	0.042	0.042
	30	0.049	0.049	0.049	0.044	0.044
	40	0.048	0.048	0.048	0.038	0.038
	50	0.058	0.058	0.058	0.068	0.068
	100	0.050	0.050	0.050	0.053	0.053
	200	0.053	0.053	0.053	0.056	0.056
3	20	0.059	0.055	0.059	0.050	0.050
	30	0.054	0.051	0.050	0.045	0.039
	40	0.052	0.050	0.050	0.051	0.047
	50	0.053	0.053	0.054	0.056	0.052
	100	0.052	0.051	0.052	0.048	0.051
	200	0.051	0.052	0.053	0.049	0.052
4	20	0.058	0.049	0.051	0.050	0.045
	30	0.053	0.051	0.049	0.056	0.048
	40	0.049	0.052	0.051	0.042	0.051
	50	0.053	0.054	0.050	0.057	0.050
	100	0.055	0.053	0.056	0.057	0.050
	200	0.050	0.049	0.050	0.053	0.051
5	20	0.050	0.056	0.055	0.049	0.046
	30	0.050	0.054	0.054	0.056	0.050
	40	0.052	0.051	0.051	0.056	0.048
	50	0.049	0.050	0.050	0.049	0.045
	100	0.050	0.050	0.050	0.049	0.048
	200	0.047	0.050	0.053	0.051	0.052
Mean*		0.052	0.052	0.052	0.051	0.049
Median*		0.052	0.051	0.051	0.051	0.050
SD*		0.003	0.002	0.003	0.006	0.006

*Column Mean, Median, and SD.

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Table 4(a): Comparison of agreement measures under the alternative situations: Sample Mean from 10,000 simulations: $K = 3$.

Configuration	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
1	20	0.388	0.435	0.482	0.760	0.840
	30	0.390	0.438	0.488	0.760	0.840
	40	0.393	0.442	0.492	0.760	0.841
	50	0.394	0.442	0.493	0.760	0.840
2	20	0.055	0.030	0.013	0.260	0.310
	30	0.053	0.027	0.010	0.253	0.302
	40	0.053	0.027	0.009	0.251	0.301
	50	0.053	0.026	0.009	0.251	0.301
3	20	0.169	0.189	0.206	0.798	0.838
	30	0.164	0.184	0.202	0.800	0.840
	40	0.168	0.189	0.207	0.799	0.839
	50	0.170	0.192	0.211	0.799	0.840
4	20	0.002	0.002	0.002	0.700	0.780
	30	0.000	-0.001	-0.001	0.699	0.779
	40	-0.001	-0.002	-0.003	0.699	0.779
	50	0.001	0.002	0.002	0.700	0.780
5	20	0.359	0.359	0.359	0.876	0.938
	30	0.360	0.360	0.360	0.875	0.937
	40	0.364	0.364	0.364	0.875	0.938
	50	0.367	0.367	0.367	0.875	0.938
6	20	-0.280	-0.362	-0.434	0.350	0.451
	30	-0.284	-0.369	-0.443	0.350	0.451
	40	-0.287	-0.373	-0.448	0.350	0.449
	50	-0.288	-0.376	-0.453	0.350	0.450

Table 4(b): Comparison of agreement measures under the alternative situations: Statistical Power from 10,000 simulations: $K = 3$.

Configuration	N	κ	$\kappa_{w^{(1)}}$	$\kappa_{w^{(2)}}$	AI_1	AI_2
1	20	0.708	0.735	0.641	0.755	0.585
	30	0.844	0.883	0.817	0.876	0.777
	40	0.934	0.955	0.914	0.963	0.904
	50	0.971	0.983	0.958	0.987	0.962
2	20	0.172	0.001	0.018	0.936	0.981
	30	0.236	0.000	0.017	0.994	0.998
	40	0.313	0.002	0.011	0.999	1.000
	50	0.384	0.003	0.011	1.000	1.000
3	20	0.246	0.243	0.236	0.861	0.559
	30	0.269	0.273	0.268	0.950	0.740
	40	0.319	0.331	0.309	0.988	0.861
	50	0.379	0.386	0.364	0.997	0.935
4	20	0.049	0.051	0.049	0.450	0.259
	30	0.047	0.049	0.046	0.560	0.357
	40	0.044	0.045	0.045	0.713	0.473
	50	0.047	0.051	0.046	0.814	0.620
5	20	0.402	0.402	0.402	0.999	1.000
	30	0.540	0.540	0.540	1.000	1.000
	40	0.650	0.650	0.650	1.000	1.000
	50	0.740	0.740	0.740	1.000	1.000
6	20	0.475	0.597	0.586	0.686	0.727
	30	0.664	0.786	0.771	0.876	0.851
	40	0.789	0.894	0.883	0.952	0.942
	50	0.883	0.949	0.941	0.978	0.970

WEIGHTS FOR WEIGHTED KAPPA AS A MEASURE OF AGREEMENT

Table 4(c): Comparison of agreement measures under the alternative situations: Sample Variance and Bias from 10,000 simulations: $K = 3$.

Conf.	N	Var		Var		Var		Var		Var	
		$Var(\kappa)$	%bias	$(\kappa_{w^{(1)}})$	%bias	$(\kappa_{w^{(2)}})$	%bias	(AI_1)	%bias	(AI_2)	%bias
1	20	0.028	14.1	0.029	-1.0	0.038	-18.8	0.005	-23.8	0.004	-46.2
	30	0.019	14.7	0.019	-2.5	0.025	-21.6	0.003	-23.3	0.003	-46.0
	40	0.013	8.9	0.014	-8.0	0.018	-25.7	0.002	-27.2	0.002	-48.6
	50	0.011	10.5	0.011	-6.6	0.015	-24.4	0.002	-25.2	0.001	-47.0
2	20	0.003	1.8	0.001	-58.9	0.001	-42.2	0.006	-4.8	0.008	17.9
	30	0.002	5.6	0.000	-60.4	0.001	-42.5	0.004	-2.6	0.006	19.7
	40	0.001	7.5	0.000	-59.0	0.001	-41.4	0.003	0.2	0.004	23.0
	50	0.001	7.2	0.000	-59.4	0.000	-42.0	0.003	2.0	0.003	25.7
3	20	0.042	62.8	0.050	56.5	0.065	54.8	0.006	-11.1	0.005	-23.7
	30	0.027	49.0	0.032	46.0	0.043	47.6	0.004	-12.9	0.003	-26.0
	40	0.020	44.4	0.024	39.5	0.032	40.1	0.003	-13.0	0.003	-26.0
	50	0.016	46.4	0.020	42.0	0.026	41.9	0.002	-9.8	0.002	-23.0
4	20	0.029	6.7	0.033	7.4	0.046	7.8	0.007	-3.7	0.006	-18.7
	30	0.019	2.8	0.022	3.1	0.031	3.5	0.004	-5.3	0.004	-20.9
	40	0.014	-1.1	0.016	-0.5	0.023	0.5	0.003	-4.7	0.003	-18.9
	50	0.012	1.9	0.013	2.2	0.019	2.8	0.003	-4.3	0.002	-19.3
5	20	0.054	16.5	0.054	16.5	0.054	16.5	0.002	-66.2	0.001	-91.7
	30	0.035	11.7	0.035	11.7	0.035	11.7	0.002	-66.2	0.000	-91.7
	40	0.027	10.7	0.027	10.7	0.027	10.7	0.001	-65.7	0.000	-91.6
	50	0.021	10.2	0.021	10.2	0.021	10.2	0.001	-66.1	0.000	-91.7
6	20	0.017	-25.3	0.022	-26.9	0.033	-24.8	0.006	-5.3	0.009	24.4
	30	0.011	-29.3	0.014	-31.0	0.022	-28.3	0.004	-6.3	0.006	25.4
	40	0.008	-29.1	0.011	-32.0	0.016	-30.5	0.003	-6.0	0.004	24.7
	50	0.007	-31.1	0.009	-32.6	0.013	-30.5	0.003	-4.6	0.004	26.9

Using Exploratory Factor Analysis for Locating Invariant Referents in Factor Invariance Studies

W. Holmes Finch
Ball State University

Brian F. French
Washington State University

Model identification in multi-group confirmatory factor analysis (MCFA) requires an equality constraint of referent variables across groups. Invariance assumption violations make it difficult to locate parameters that actually differ. Suggested procedures for locating invariant referents are cumbersome, complex, and provide imperfect results. Exploratory factor analysis (EFA) may be an alternative because of its ease of use, yet empirical evaluation of its effectiveness is lacking. EFAs accuracy for distinguishing invariant from non-invariant referents was examined.

Key words: Factor analysis, invariance, bias.

Introduction

The use of scores for making decisions about persons, be it for job placement, graduation from high school, acceptance to graduate school, or obtaining a license to operate a motor vehicle, relies on the continued accumulation of empirical and theoretical validity evidence to support such score use (Messick, 1989). One form of empirical validity evidence is measurement invariance or equivalence. An assessment instrument, for example, should have the same psychometric properties across groups to help ensure that measurement of the specified construct is the same across groups. In the absence of such evidence, group comparisons on the ability of interest may be meaningless, as observed differences could be the result of ability differences or measurement differences (i.e., a lack of invariance).

Holmes Finch is an Assistant Professor of research methods and statistics in the Teachers College at Ball State University. E-mail: frenchb@purdue.edu. Brian French is an Associate Professor teaching courses in research methods and statistics in the College of statistics in the College of Education at Washington State University. Email: frenchb@wsu.edu

A common method for assessing factor invariance, a form of measurement invariance (MI), is multi-group confirmatory factor analysis (MCFA). MCFA allows for an *a priori* specified latent structure of an instrument to be assessed for MI across groups or time (e.g., Alwin & Jackson, 1981; Golembiewski, Billingsley, & Yeager, 1976). A powerful feature of MCFA is the ability to compare specific model features (e.g., factor loadings) at the matrix level, as well as individual elements of the matrix under examination.

Invariance testing in MCFA involves comparing increasingly more restricted factor models by sequentially constraining different parameter estimates (e.g., factor loadings, error variances) invariant or equal across groups or time. The presence of MI is determined using differences in the chi-square goodness-of-fit statistics for more and less restrictive models, where a non-significant difference indicates invariance. This procedure has been well documented both in theoretical and applied examples (e.g., Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1996; Maller & French, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993).

There are several procedural aspects of invariance testing that deserve further attention before practitioners and researchers have complete confidence in such results (Little, 2000). One of several unresolved issues in MCFA is the need to constrain a referent

indicator equal across groups (Millsap, 2005). Latent factors are constructed on arbitrary coordinate systems making comparison of models across populations difficult because they are not constrained to the same system in relation to the other populations or groups of interest (Wilson, 1981).

The model standardization, or identification, procedure can solve this problem by assigning units of measurement to the latent variables (Jöreskog & Sörbom, 1996), generally by aligning the latent factors to a scale based on the same indicators across groups. To meet identification requirements, per factor, either a factor variance or a factor loading is set to 1.0 across groups. Additional methods have been suggested (e.g., see Drasgow & Kanfer, 1985; Reise, Widaman, & Pugh, 1993), but the factor loading method appears to be used most commonly (Brown, 2006; Vandenberg & Lance, 2000). These procedures require the assumption that the referent variable constrained equal is, in fact, invariant. This assumption cannot be directly tested, however, because only the ratio of individual factor loadings to the referent can be compared across groups (Bielby, 1986; Cheung & Rensvold, 1999; Wilson, 1981). Furthermore, complications arise as different constraint choices may lead to different results in terms of model fit and hypotheses concerning equality of parameters (Millsap, 2001; Steiger, 2002; Wilson).

When the referent parameter is not invariant, estimates of other parameters may be distorted, which can lead to inaccurate conclusions regarding their invariance (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2005). A circular situation exists with this assumption where (a) the referent variable must be invariant, (b) invariance cannot be established without estimating a model, and (c) model estimation requires an invariant referent. Thus, we are back to the original invariant referent assumption. That is, to assess invariance for a given factor loading across groups, for instance, an equality constraint (that is actually true) must already be placed on another factor loading. This circular conundrum is parallel to the ability purification process in detection of differential item functioning (DIF) (e.g., Holland & Thayer, 1988; Lord, 1980), another method commonly

employed to establish MI at the item level. Ability purification in DIF analysis attempts to identify a set of non-DIF items for use as the matching criterion and can lead to more accurate DIF detection (Ackerman, 1992; Clauser et al., 1993). A similar procedure with MCFA would seem appropriate with the expected outcome of more accurate detection of a lack of MI.

A search procedure (i.e., factor-ratio test and the stepwise partitioning procedure) was designed to identify invariant and non-invariant variables (Rensvold & Cheung, 2001). The method uses each variable, in turn, as the referent in a set of models with each other variable constrained to be invariant. The iterative procedure tests all pairs of variables (i.e., $p(p-1)/2$ pairs) and becomes quite complex with many indicators, making it not “user-friendly” for practitioners (Vandenberg, 2002). A moderate length instrument (i.e., 30 indicators), for instance, requires 435 individual invariance tests. Furthermore, empirical evaluation of the method demonstrated adequate (e.g., acceptable false and true positives) but far from perfect performance (French & Finch, 2006a).

To overcome these limitations, exploratory factor analysis (EFA) may be a viable alternative for identifying invariant referents, as a purification step prior to a MCFA. That is, if a researcher intends to set one loading invariant across groups, a single EFA could be conducted for each group separately and loading estimates compared to ascertain which loadings appear to be invariant. With an EFA conducted on each group separately, such an analysis may be considered a weak test of factorial invariance (Zumbo, 2003).

EFA is not a formal test of invariance, but instead is a possible method to examine parameter estimates across groups to obtain a sense of the differences in the factor loadings without need of conducting a large number of analyses as is required when using the factor-ratio test. Specifically, pattern coefficients appearing most similar would be eligible for serving as a referent variable in the MCFA. Such use is in accord with suggestions that EFA be used to examine loadings with an “interocular eyeball test” (Vandenberg, 2002) to judge similarity of loadings to identify appropriate

referent variables. However, there does not appear to be empirical evaluation of EFA for locating potentially invariant referents.

The purpose of this study is to assess the utility of EFA in identifying non-invariant or invariant factor loadings between two groups. This procedure would be used prior to the actual MCFA as a “purification” process for identifying a loading that is likely to be invariant for use as the referent parameter. The procedure would simply entail conducting one EFA per group with one reference group and one comparison group. The loadings (i.e., pattern coefficients) from the separate analyses would be compared visually to determine similarity of individual loadings. Loadings that appear markedly different would not be used as a referent, while loadings appearing most similar would be used as the referent. If multiple loadings across groups were equally similar, any of them could serve as the referent.

Methodology

Simulated data were used to control variables that could influence the magnitude of factor loading estimates, with 1000 replications for each combination of conditions described below. Simulations and analyses were completed in SAS, V9.1 (The SAS Institute, 2003).

Number of Factors and Indicators

Data were simulated from both 1- and 2-factor models, with interfactor correlations set at .50 to represent moderately correlated factors. The number of indicators per factor was 6. Data were simulated to reflect simple structure for continuous and normally distributed subtest level data.

Sample Size

The necessary sample size to obtain reasonable estimates in factor analysis varies depending on the data conditions. For this reason, three sample size conditions were simulated: 100, 500, and 1000 in order to reflect small, medium and large samples. These values are consistent with other factor analysis simulation studies (e.g., Cheung & Rensvold, 2002; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004), ranging from poor ($n=$

100) to excellent ($n=1000$) (Comery & Lee, 1992), and may not be of much concern here as communalities were high (MacCallum, Widaman, Zhang, & Hong, 1999).

Magnitude of Difference with the Non-Invariant Indicators

Six levels of factor loading values for the non-invariant indicator were simulated. A baseline condition was established where no differences in loadings were present, so that the first indicator had a loading value of 0.75, as did the other variables. The remaining 5 conditions were characterized by declines in the target loading from 0.10 to 0.50 in increments of 0.10 (i.e., 0.65, 0.55, 0.45, 0.35, and 0.25). These levels were selected as there is no effect size, at least to the knowledge of the authors, for what represents a meaningful factor loading difference (Millsap, 2005) and the range covers values used in previous MCFA simulation work (e.g., French & Finch, 2006b; Meade & Lautenschlager, 2004).

Contamination

The location of invariant parameters may be influenced by the number of indicators that lack invariance (Millsap, 2005). Thus, the presence of a factor loading exhibiting a difference from 0.75 other than that for the target indicator was varied as either present or absent. In other words, for half of the simulated conditions only the target indicator loading was contaminated, while for the other half of the simulations a second target indicator loading also was contaminated at the same difference as the target indicator. This allowed assessment of the influence of additional contaminated variables.

Analysis

All analyses were conducted by group using principal axis factoring with PROMAX rotation in the 2-factor condition. These settings follow recommendations for using EFA for a referent indicator search and are more consistent with educational and psychological data (e.g., presence of measurement error, correlated factors; Vandenberg, 2002).

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Evaluation Criteria

Factor loadings (i.e., pattern coefficients) obtained from the EFA for the target variable were compared with 0.75, which was the population value for the reference group. The assumption of this study was that if a researcher were to use EFA to identify invariant indicators, the observed loadings would be compared between the two groups, as described above. Therefore, performance could be judged by how well EFA would estimate factor loading values for the second group both when they differ in the population from that of the reference group, and when they do not. Three statistics across replications were used to operationalize this overall outcome: (a) the mean loading for the target variable (loading bias), (b) the standard deviation of the target loading, and (c) the percent of replications for which the observed loading was within 10% of the baseline loading of 0.75; i.e. between 0.675 and 0.825. This latter criterion was selected because of suggestions that bias values less than 10-15% may not be considered serious in many latent variable modeling situations (Muthén, Kaplan, & Hollis, 1987).

Results

Factor loading bias

Based on the Analysis of variance ANOVA ($\alpha = 0.05$) used to identify the manipulated variables and their interactions that were associated with factor loading bias, the 3-way interaction of magnitude of difference by number of factors by contamination was the highest order significant term. Other 2-way interactions involving combinations of these three variables also were statistically significant, as were the main effects of number of factors and the magnitude of the difference. The 3-way interaction had an η^2 value of only 0.02, while the magnitude of loading difference had an η^2 of 0.94. Thus, while the interaction should not be ignored, it is clear that the most important factor in determining the mean loading is the magnitude of the difference from the baseline of 0.75. For this reason, both terms are discussed below.

The means of factor loading estimates across the magnitude of difference, number of

factors and level of contamination appear in Table 1.

These values demonstrate that EFA, using principal axis factoring and PROMAX rotation, accurately estimates the population factor loading of 0.75 for both the 1- and 2-factor conditions when all other loadings also are 0.75. Furthermore, the estimates also were very close to the population value of 0.75 when a loading other than that for the target variable was set at 0.65 (i.e., contaminated condition). When the target loading was different from 0.75 in the population, the sample estimate was generally very close to the actual population value in the 1-factor case, regardless of whether other factor loadings were contaminated. This result was mirrored in the 2-factor case with no contamination. However, when non-target loadings were contaminated, the means of the target loadings reflect overestimation except when the target was 0.65. As expected due to high communalities, sample size was not significantly related to the mean value of the estimated factor loadings.

Standard Deviation

The ANOVA identified the interaction of sample size by number of factors by magnitude of the difference as statistically significant for the standard deviation of loading estimates. In addition, the main effects of magnitude of difference, sample size, number of factors, and contamination were also statistically significant. It should be noted that the sample size accounted for 75.5% of the variation in the standard deviation, while none of the other terms in the model accounted for more than 4%.

Table 2 contains the standard deviations of the factor loading estimates by the number of factors, sample size, and magnitude of difference between the target loading and 0.75. An examination of these results suggests that in general, larger sample sizes were associated with lower variation in the estimates. In addition, as the magnitude of the difference increased, the standard deviation did as well. This effect was slightly more pronounced for smaller samples. Finally, the difference in standard deviations by sample size was slightly greater in the 2-factor case. Again, it is important to note that while this interaction was found to be statistically

Table 1. Mean of Factor Loadings across Replications by Number of Factors, Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
0.75	0.742	0.741
0.65	0.647	0.645
0.55	0.551	0.550
0.45	0.453	0.454
0.35	0.353	0.356
0.25	0.255	0.271
	<u>2 Factor</u>	
0.75	0.735	0.731
0.65	0.641	0.580
0.55	0.548	0.578
0.45	0.451	0.490
0.35	0.353	0.390
0.25	0.255	0.339

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 2. Standard deviation of Factor Loadings across Replications by Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Sample size</u>		
<u>Factor 1</u>	100	500	1000
0.75	0.049	0.022	0.016
0.65	0.065	0.029	0.021
0.55	0.076	0.034	0.024
0.45	0.089	0.038	0.028
0.35	0.097	0.043	0.029
0.25	0.105	0.047	0.033
<u>Factor 2</u>			
0.75	0.062	0.028	0.026
0.65	0.084	0.037	0.023
0.55	0.092	0.041	0.029
0.45	0.103	0.046	0.033
0.35	0.114	0.052	0.036
0.25	0.147	0.052	0.037

significant, it accounted for less than 5% of the variance, whereas sample size accounted for 75% of the variance in the standard deviation values.

Logistic regression was used to identify significant main effects and interactions that were associated with loadings being within a 10% range of 0.75. In this case, the outcome for each replication was coded as either 1 (loading was within 10% of 0.75) or 0 (loading was not within this range). It is important to keep in mind that when the target loading for the second group was simulated to be 0.75, the proportion

of cases not within this 10% range could be considered a false positive (incorrect identification of difference when no difference existed). On the other hand, when the target loading was simulated to be some other value (e.g., 0.25), the proportion of cases outside of the 10% range represent a true positive (i.e., correct identification of differences between the groups' loadings). Two separate logistic regression models were used: 1) Examining only those cases where the target loading was set at 0.75 (Model 1) and 2) Examining all other target loading conditions (Model 2).

Table 3. Proportion of Factor Loadings within 10% of 0.75 (0.675, 0.825) when loading was 0.75 across Replications by Number of Factors and Sample Size

<u>Sample size</u>	<u>Factors</u>	
	<u>1</u>	<u>2</u>
100	0.875	0.771
500	0.998	0.979
1000	1.000	0.996

In the case of Model 1, the only statistically significant effects were the interaction of the number of factors by sample size and the main effect of sample size. Table 3 contains the proportion of cases within 10% of 0.75 when the target loading was in fact 0.75 in the population, by sample size and the number of factors. The results show that for the 1-factor case, over 87% of the factor loading estimates were within the correct range, regardless of sample size. In contrast, for the 2-factor case, the smallest sample size was associated with a somewhat lower proportion of cases within the 10% range of the 0.75 value compared to the 1-factor case, otherwise the results across factor models were nearly identical.

In the case of Model 2, the logistic regression analysis identified the 3-way interactions of number of factors by contamination by magnitude of difference and number of factors by sample size by contamination as significantly associated with the proportion of cases within 10% of the 0.75 loading value. In this context being outside of this range would be correct, given that the population values for the simulated loadings were less than 0.75.

Table 4 contains the proportion of replications within 10% of 0.75 by the magnitude of the difference, the number of factors and contamination. For the 1-factor case, regardless of contamination, the larger the target loading was in the population (i.e., less of a

difference), the greater the proportion of replications for which the estimated value was within 10% of 0.75. The largest proportion of values within this range occurred for the population loading of 0.65 across the number of factors and level of contamination.

Indeed, the results for the 1-factor cases (both contaminated and not) and the 2-factor uncontaminated case were all very comparable. However, in the contaminated 2-factor condition with a population loading of 0.65, the proportion of replications within 10% of 0.75 (i.e., 0.055) was much lower than in the other 3 conditions ($M = 0.192$). For the other loading values, the results for the contaminated 2-factor case were just slightly higher than for the others simulated.

Table 5 displays the proportion of replications within 10% of 0.75 by the number of factors, sample size and contamination condition. Overall, the proportions decline in conjunction with increasing sample sizes. For 1-factor these proportions were very comparable regardless of whether another loading was different from 0.75 (contaminated condition). While the pattern of changes in the proportion declined with increasing sample sizes in the 2-factor case, there was a slightly greater difference in the proportions between the contaminated and uncontaminated conditions, leading to the significant interaction described above.

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 4. Proportions of Factor Loadings within 10% of 0.75 (0.675, 0.825) across Replications by Number of Factors, Sample Size and Population Target Loading Value

<u>Loading for Group 2</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
0.65	0.194	0.185
0.55	0.013	0.015
0.45	0	0.001
0.35	0	0
0.25	0	0
	<u>2 Factor</u>	
0.65	0.197	0.055
0.55	0.031	0.047
0.45	0.002	0.011
0.35	0	0.001
0.25	0	0.001

Discussion

The results reported in this study suggest that in many instances EFA may be a useful tool for identifying potential indicator variables with invariant loadings across groups for use in a subsequent MCFA. Across most of the conditions simulated here, the factor loading estimates provided by principal axis EFA with PROMAX rotation were very close to the population values. Indeed, the only instances where simulated values were not approximated occurred with 2 factors in conjunction with the contamination of one other factor loading. These generally positive results would seem to suggest that practitioners using EFA can be confident that the sample estimate of loadings are unbiased

in conditions such as those simulated here.

The amount of variation in sample estimates was largely a function of sample size. While loading estimates had greater variability across replications for smaller loading values in the population, there were more marked differences in variation across the three sample size conditions. In addition, this difference in variability was largely mitigated by sample size, so that for 100 participants the standard deviation increased by as much as 0.8 (2 factors) as the population loading value declined, while for 500 or 1000 participants, this increase was always less than 0.03.

In short, with sufficient sample size, a researcher using EFA to identify invariant factor loadings can be almost as confident in their result whether the loading is at or near 0.75 or

closer to 0.25. Note that a certain sample size (e.g., $N = 500$) requirement is not being recommended, but rather that the sample size be sufficient given the data conditions (e.g., number of variables, communalities), as sample sizes requirements for accurate estimate can depend on data conditions (e.g., MacCallum et al., 1999).

In terms of the identification of false positives (obtaining a sample estimate that was more than 10% different than 0.75 when it was not in the population), EFA appears to have performed better for larger sample sizes, particularly in the 2-factor case. Indeed, with a sample size of 500 or greater with the population loading set at 0.75, the likelihood of making a false positive was essentially 0.02 or less. That is, the sample estimate was within the expected range 98 % of the time or higher. In contrast, the rate of sample estimates being within 10% of 0.75 when they should not have been (i.e. the population loading was not 0.75) declined as (a) the value of the population loading declined increasingly from 0.75, and (b) as sample size increased. Given that the results have shown generally little or no bias in loading estimates, this outcome is not a surprise. Indeed, if the target loading was 0.55 or lower in the population, the sample estimates were within 10% of 0.75 in fewer than 5% of cases, regardless of contamination condition. Thus, supporting that EFA could quite accurately detect a non-invariant loading.

The identification of an invariant referent loading is a crucial step in MCFA. As described above, a failure to accurately select an invariant parameter value in the model identification step could lead to severely biased parameter estimates (e.g., factor loadings) which in turn could compromise other analyses, such as the comparison of latent means. The primary method suggested in the literature for identifying invariant indicators, or sets of indicators, is the factor-ratio test and SP procedure (Rensvold & Cheung, 2001), which involves a fairly complex and time consuming multi-step analysis. While this approach appears to work reasonably well for fairly limited models it can become intractably time consuming with increasing model complexity (French & Finch, 2006a).

EFA is one approach that has been advocated for use in practice and involves comparison of factor loading estimates between two groups (Vandenberg, 2001; Zumbo, 2003). While this method does not have the advantage of significance testing that is offered by the factor-ratio test, it is much simpler to conduct. The results of this study seem to indicate that in conditions such as those simulated here, EFA generally provides unbiased estimates of factor loadings, which can in turn be compared to a target value (such as those of another group in the MCFA context).

Therefore, practitioners interested in identifying loadings that are invariant across groups may find that this simple approach works quite well in conditions similar to those simulated here. It does seem that greater confidence can be placed in EFA factor loading estimates that are based on larger sample sizes, particularly with respect to false negative outcomes when the population loadings for the groups differ by 0.10 or more. Under such conditions, the EFA approach appears to have low false negative rates (below 0.05). In addition, the lack of bias and the lower standard deviations at sample sizes of 500 or more appear to contribute to the ability of EFA to accurately estimate loadings within 10% of the target value. Study limitations and directions for future research

As with many simulation studies, the generalizability of the results is limited due to the conditions under study, which should be remembered when interpreting these results. First, the factor models simulated were not as complex as seen in some invariance studies. While the EFA worked well for these somewhat simpler models, it will be necessary to assess its performance with more complex problems (e.g., greater number of factors, different variables, various levels of communalities). Second, a related area that deserves attention is the combination of loadings for the observed variables. In this study, all of the loadings were set at 0.75 (unless contaminated). Given that this is one of the first (if not the first) Monte Carlo investigations to examine the use of EFA to accurately identify invariant referent variables, clarity of result interpretation was considered

INVARIANT REFERENTS IN FACTOR INVARIANCE STUDIES

Table 5. Proportion of Factor Loadings within 10% of 0.75 (0.675, 0.825) across Replications by Number of Factors, Sample Size, and Contamination

<u>Sample size</u>	<u>Contaminated</u>	
<u>1 Factor</u>	<u>No</u>	<u>Yes</u>
100	0.079	0.080
500	0.030	0.027
1000	0.015	0.013
	<u>2 Factor</u>	
100	0.090	0.066
500	0.033	0.002
1000	0.014	0.000

paramount, and thus non-target loadings were not varied. However, it is unclear whether the results obtained here would hold for a more complex combination of loading values and factor models, as well as data conditions (e.g., ordinal variables). Thus, although EFA appears to be a promising screening or purification tool prior to MCFA analysis, future research should extend the current work by investigating a broader combination of conditions before the tool is applied unequivocally.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, 29, 67-91.

Alwin, D. F. & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249-279). Beverly Hills: Sage.

Bielby, W. T. (1986). Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods & Research*, 15, 3-23.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. New York: The Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *Psychological Bulletin*, 105, 456-466.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, 25, 1-27.

Clauser, B., Mazor, K., & Hambleton, R. K. (1993). The effects of purification of the matching criterion on the identification of DIF using the Mantel-Haenszel procedure. *Applied Measurement in Education*, 6, 269-279.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*. Hillsdale, N.J.: Erlbaum.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology*, 70, 662-680.

French, B. F., & Finch, W. H. (2006a, June). *Locating the Invariant Referent in Multi-Group Confirmatory Factor Analysis*. Paper presented at the International Psychometric Society meeting in Montreal, Canada.

French, B. F., & Finch, W. (2006b). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, 13, 378-402.

Golembiewski, R.T., Billingsley, K. & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science*, 12, 133-157.

- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Holland & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.
- Jöreskog, K.G. & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Chicago: ScientificSoftware.
- Little, T. D. (2000). On the comparability of constructs in cross-cultural research: A critique of Cheung and Rensvold, *Journal of Cross-Cultural Psychology*, 31, 213-219.
- Lubke, G. H. & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, 11, 514-534.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.
- MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, 4, 84-99.
- Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement*, 64, 647-660.
- Meade, A. W., & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, 11, 60-72.
- Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, 18, 5-11.
- Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, 8, 1-17.
- Millsap, R.E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J.J. McArdle (Eds.) *Contemporary Psychometrics* (pp. 153-172). Mahwah, NJ: Lawrence Erlbaum Associates.
- Muthén, B., Kaplan, D. & Hollis, M (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431-462.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology*, 87, 517-529.
- Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches to exploring measurement invariance. *Psychological Bulletin*, 114, 552-566.
- Rensvold, R. B., & Cheung, G. W. (1998). Testing measurement models for factorial invariance: A systematic approach. *Educational and Psychological Measurement*, 58, 1017-1034.
- Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Equivalence in measurement* (pp. 25-50). Greenwich, CT: Information Age Publishing.
- SAS Institute (2004) *SAS version 9.1.3*. Cary, NC: SAS Institute inc.
- Steiger, J. H. (2002). When constraints interact: A caution about reference variables, identification constraints, and scale dependencies in structural equation modeling. *Psychological Methods*, 7, 210-227.
- Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, 5, 139-158.
- 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, 4-69.
- Wilson, K. L. (1981). On population comparisons using factor indexes or latent variables. *Social Science Research*, 10, 301-313.
- Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing*, 20, 136-147.

Using Connectionist Models to Evaluate Examinees' Response Patterns to Achievement Tests

Mark J. Gierl Ying Cui Steve Hunka
University of Alberta

The attribute hierarchy method (AHM) applied to assessment engineering is described. It is a psychometric method for classifying examinees' test item responses into a set of attribute mastery patterns associated with different components in a cognitive model of task performance. Attribute probabilities, computed using a neural network, can be estimated for each examinee thereby providing specific information about the examinee's attribute-mastery level. The pattern recognition approach described in this study relies on an explicit cognitive model to produce the expected response patterns. The expected response patterns serve as the input to the neural network. The model also yields the cognitive test specifications. These specifications identify the examinees' attribute patterns which are used as output for the neural network. The purpose of the statistical pattern recognition analysis is to estimate the probability that an examinee possess specific attribute combinations based on their observed item response patterns. Two examples using student response data from a sample of algebra items on the SAT illustrate our pattern recognition approach.

Keywords: Attribute hierarchy method, multilayer perceptron, neural network, educational measurement.

Introduction

Educational measurement is undergoing profound changes, as developments in cognitive science, mathematical statistics, computer tech-

nology, educational psychology, and computing science are permeating the testing field. In particular, the influence of cognitive psychology on educational measurement, which began almost 20 years ago (Snow & Lohman, 1989), has become a source of great activity contributing to many of the ideas and innovations in cognitive diagnostic assessment (Leighton & Gierl, 2007a). One consequence of these interdisciplinary influences is the emergence of a new area of research called assessment engineering (AE) (Luecht, 2006). AE is an innovative approach to measurement where engineering-like principles are used to direct the design as well as the analysis, scoring, and reporting of assessment results. With this approach, an assessment begins with specific, empirically-derived cognitive models of task performance. Next, assessment task templates are created using established frameworks derived from the cognitive model to produce replicable test items. Finally, psychometric models are applied to the examinee response data collected using the templates to produce scores that are both replicable and interpretable.

Mark J. Gierl is a Professor of Educational Psychology and Canada Research Chair in Educational Measurement, Centre for Research in Applied Measurement and Evaluation. His primary research interests include cognitive diagnostic assessment, and assessment engineering. E-mail: mark.gierl@ualberta.ca. Ying Cui is an Assistant Professor of Educational Psychology. Her primary research interests include investigating cognitive factors underlying student performance and using cognitive models to interpret student test performance. E-mail her at yc@ualberta.ca. Stephen Hunka is Professor Emeritus. His areas of specialization include statistical models and computation. E-mail him at: Albertasteve.hunka@ualberta.ca.

AE differs from more traditional approaches to test design and analysis in four fundamental ways. First, cognitive models guide task design and item development, rather than content-based test specifications. While the database “tags” associated with content specifications can be included in the task templates, the assessment principles used to develop items are much more specific allowing items to be created quickly and efficiently during the development cycle. Second, explicit data models and assessment task templates are created to control and manipulate both the content and cognitive attributes of the items. Item writers are required to use the templates during development thereby producing items that adhere to strict quality controls and that meet high psychometric standards. Third, automated test assembly procedures are employed to build assessments that function to exacting specifications, as outlined in the task templates. Hence, multiple test forms can be created from a bank of items very efficiently according to both content and statistical specifications. Fourth, pursuant to scoring and score-reporting, psychometric models are employed in a confirmatory—versus exploratory—manner to assess the model-data fit relative to the intended underlying structure of the constructs or traits the test is design to measure. The outcomes from these model-data fit analyses also provide developers with guidelines for specific modifications to the cognitive models and task templates, as needed, to facilitate the acquisition of data that supports the intended assessment inferences.

Overview of Attribute Hierarchy Model

Recently, Leighton, Gierl, and Hunka (2004; see also Gierl, Leighton, & Hunka, 2007) proposed the attribute hierarchy method (AHM). The AHM is a psychometric method used to classify examinees’ test item responses into a set of structured attribute patterns associated with different components from a cognitive model of task performance (Leighton & Gierl, 2007b). Attributes include different procedures, skills, and/or processes that an examinee must possess to solve a test item. These attributes are structured using a hierarchy so the ordering of the cognitive skills is specified. As a result, the

attribute hierarchy serves as an explicit cognitive model. This model, in turn, provides the structure for both developing test items and linking examinees’ test performance to specific cognitive inferences about skill acquisition. The AHM was developed to address two specific problems associated with *feature creation* and *statistical pattern recognition* (Gierl, 2007). Our solutions to these problems are described in the next two sections.

Feature Creation with the AHM

To make specific inferences about problem solving, cognitive models are required to operationalize the construct of interest. A cognitive model in educational measurement refers to a simplified description of human problem solving on standardized tasks at some convenient grain size or level of analysis in order to facilitate explanation and prediction of students’ performance. These models provide an interpretative framework that can guide item development so test performance can be linked to specific cognitive inferences about examinees’ knowledge, processes, and strategies. These models also provide the means for connecting cognitive principles with measurement practices.

A cognitive model of task performance is specified at a small grain size because it magnifies the cognitive processes underlying test performance. Often, a cognitive model of task performance will also reflect a hierarchy of cognitive processes within a domain because cognitive processes share dependencies and function within a much larger network of inter-related processes, competencies, and skills. Assessments based on cognitive models of task performance should be developed so test items directly measure specific cognitive processes of increasing complexity in the examinees’ understanding of a domain. The items can also be designed with this hierarchical order in mind, so that test performance is directly linked to information about students’ cognitive strengths and weaknesses. Strong inferences about examinees’ cognitive skills can be made because the small grain size in these models help illuminate the knowledge and skills required to perform competently on testing tasks. Specific diagnostic inferences can also be generated

when items are developed to measure different components and processes in the model.

To specify the relationships among the attributes in the hierarchy using the AHM, the adjacency and reachability matrices are defined. The direct relationship among attributes is specified by a binary adjacency matrix (A) of order (k, k) , where k is the number of attributes, such that the ij^{th} element represents the absence (i.e., 0) or presence (i.e., 1) of a direct connection between two attributes. The adjacency matrix is of upper triangular form. The direct and indirect relationships among attributes are specified by the binary reachability matrix (R) of order (k, k) , where k is the number of attributes. To obtain the R matrix from the A matrix, Boolean addition and multiplication operations are performed on the adjacency matrix, meaning $R = (A + I)^n$, where n is the integer required to reach invariance, $n = 1, 2, \dots, m$, and I is the identity matrix.

Next, the potential pool of items is generated. This pool is considered to be those items representing all combinations of attributes when the attributes are independent of one other. The size of the potential pool is $2^k - 1$, where k is the number of attributes. The attributes in the potential pool of items are described by the incidence matrix (Q) of order (k, p) , where k is the number of attributes and p is the number of potential items. This matrix can be reduced to form the reduced Q matrix (Q_r) by imposing the constraints of the attribute hierarchy as defined in the R matrix. The Q_r matrix represents the items from the potential pool that fit the constraints defined in the attribute hierarchy. The Q_r matrix is formed using Boolean inclusion by determining which columns of the R matrix are logically included in each column of the Q matrix. The Q_r matrix is of order (k, i) where k is the number of attributes and i is the reduced number of items resulting from the constraints in the hierarchy.

Given a hierarchy of attributes, the expected response patterns for a group of examinees can then be generated. The expected

response matrix (E) is created, again using Boolean inclusion, where the algorithm compares each row of the attribute pattern matrix (which is the transpose of the Q_r matrix) to the columns of the Q_r matrix. The expected response matrix, of order (j, i) , is calculated, where j is the number of examinees and i is the reduced number of items resulting from the constraints imposed by the hierarchy.

Assessment engineering principles are used explicitly with the AHM to design test items and analyze examinees' observed response patterns. To design test items, the Q_r matrix is used. Recall, the Q_r matrix is produced by determining which columns of the R matrix are logically included in columns of the Q matrix, using Boolean inclusion. The Q_r matrix can be interpreted as the cognitive test specification because it contains the attribute-by-item specification for each component of the cognitive model of task performance outlined in the A matrix. Hence, the results from the Q_r matrix can be used to develop items that measure each specific attribute combination defined in the hierarchy. Then, in the pattern recognition stage, as described in the next section, examinees' observed response patterns can be analyzed according to the cognitive characteristics probed by each item.

Pattern Recognition with the AHM

An examinee's observed response pattern is judged relative to expected response pattern with the AHM under the assumption that the cognitive model is true. Hence, the purpose of the statistical pattern recognition analysis is to estimate the probability that an examinee possess specific attribute combinations based on their response patterns. These probabilities provide examinees with specific information about their attribute-level mastery as part of the test reporting process. To estimate the probability that examinees possess specific attributes, given their observed item response pattern, an artificial neural network approach is used.

The input to train the neural network is the expected response vector derived from the cognitive model. The expected response vectors serve as the exemplars. For each expected

response vector, there is a specific combination of examinee attributes described in the transpose of the Q_r matrix. Recall, Q_r matrix is of order (k, i) where k is the number of attributes and i is the reduced number of items resulting from the constraints specified by the hierarchy. The transpose of this matrix is of order (j, k) where j is the number of examinees and k is the number of attributes. In other words, the transpose of the reduced incidence matrix has a distinct row and column interpretation—the rows serve as the examinees and the columns serve as the items. The examinee attribute patterns, like the expected response vectors, are derived from the cognitive model and, thus, specify the attribute pattern that should be associated with each expected response pattern. The relationship between the expected response vectors with their associated attribute vectors is established by presenting each pattern to the network repeatedly until it learns each association. The final result is a set of weight matrices that can be used to transform any observed response vector to its associate attribute vector. The transformed result can be interpreted as the attribute probability, scaled from 0 to 1, where a higher value indicates that the examinee has a higher probability of possessing a specific attribute (McClelland, 1998).

A multilayer perceptron is the parallel-processing architecture used in the neural network. This network transforms the stimulus received by the input unit to a signal for the output unit through the hidden units. The contribution of each input unit i to hidden unit j is determined by weight, w_{ji} . Similarly, the contribution of each hidden unit j to output unit k is determined by weight, v_{kj} . The input layer contains the exemplars (i.e., expected response patterns) the network is designed to learn. Learning is deemed to occur when the output layer, containing the desired response output (i.e., the attribute patterns), is correctly associated with the exemplars, as indicated by the value of the root mean square error. That is, the connection weights in the hidden layer

transform the input stimuli into a weighted sum defined as

$$S_j = \sum_{i=1}^p w_{ji} x_i \quad (1)$$

where S_j is the weighted sum for node j in the hidden layer, w_{ji} is the weight used by node j for input x_i , and x_i is the input from node i of the input layer with i ranging from 1 to p for the input node and j ranging from 1 to q for the hidden layer node. S_j is then transformed by the logistic function,

$$S_j^* = \frac{1}{1 + e^{-S_j}} \quad (2)$$

Similarly, the hidden layer produces a weighted linear combination of their inputs which are transformed to non-linear weighted sums that are passed to every output layer unit to produce the final attribute-level responses. The output, S_j^* , from every hidden layer unit is passed to every output layer unit where a linearly weighted sum, T_k , is formed using the weights v_{kj} , and the result transformed for output T_k^* using a nonlinear function. In other words,

$$T_k = \sum_{j=1}^q v_{kj} S_j^* \quad (3)$$

where T_k is the weighted sum for each of k output nodes using weights v_{kj} , with j ranging from 1 to q for the hidden layer nodes. T_k , like S_j , is transformed by the logistic function to T_k^* . Because the correct activation function is scaled using the logistic transformation, the output values range from 0 to 1. The result can be interpreted as the probability the correct or target value for each output will have a value of 1.

The attribute-based targets in the output units are compared to the pattern associated with the exemplars, which are the expected response patterns. However, the solution produced initially is likely to be discrepant resulting in a relatively large root mean square error. This discrepancy can be used to modify the connection weights leading to a more accurate solution and a smaller error term. With the AHM, the weights are approximated so the error term is minimized using the well-known learning algorithm called the generalized delta rule that is incorporated in the back propagation of error training procedure (Rumelhart, Hinton, & Williams, 1986a, 1986b). The final result is a set of weight matrices, one for cells in the hidden layer and one for the cells in the output layer, that can be used to transform any examinee response vector to its associate attribute vector. The functional relationship for mapping the examinees' observed response pattern onto the expected response patterns so their attribute probabilities can be computed is given as follows. Let

$$F(z) = \frac{1}{1 + e^{-z}} \quad (4)$$

and

$$a_k = \sum_{j=1}^q v_{kj} F\left(\sum_{i=1}^p w_{ji} x_i\right) \quad (5)$$

then the output for unit k , M_k^* , is given as

$$M_k^* = F(a_k) \quad (6)$$

where q is the total number of hidden units, v_{kj} is the weight of hidden unit j for output unit k , p is the total number of input units, w_{ji} is the weight of input unit i for hidden unit j , and x_i is the input received from input unit i . Using this transformation, attribute probabilities can be computed for each observed response pattern thereby providing examinees with specific information about their attribute-mastery level.

Two Examples Using SAT Algebra Items

To illustrate how a multilayer perceptron can be used to estimate the attribute probabilities in an actual testing situation, two examples are provided. Each example is based on the observed response data from a random sample of 5000 students who wrote the algebra items on the March 2005 administration of the SAT. The SAT is a college admissions test developed, analyzed, and scored by the College Board. The Mathematics section contains items in the content areas of Number and Operations; Algebra I, II, and Functions; Geometry; and Statistics, Probability, and Data Analysis. For our analysis, only a subset of items in Algebra I and II were evaluated. Sample algebra items from the SAT Mathematics section are available from the College Board website at www.collegeboard.com.

Note that cognitive models of task performance guide diagnostic inferences because they are specified at a small grain size and they magnify the cognitive processes that underlie performance. Ideally, a theory of task performance would direct the development of a cognitive model of task performance. But, in the absence of such a theory, a cognitive model must still be specified to create the attribute hierarchy. Another starting point is to develop a cognitive model from a task analysis of the items in the domain when a theory or model of task performance is unavailable. In conducting the task analysis of the SAT algebra items we, first, solved each test item and attempted to identify the mathematical concepts, operations, procedures, and strategies used to solve each item (see Gierl, Wang, & Zhou, 2006; Gierl, Leighton, Wang, Zhou, Gokiart, & Tan, 2006). These cognitive attributes were categorized so they could be ordered in a logical, hierarchical sequence to summarize problem-solving performance. The cognitive model used to characterize examinee performance on the items is presented in Figure 1. Each attribute is denoted with an A (e.g., A1, A2, etc.). Each attribute was measured by one test item. The cognitive model in Figure 1 was used to created the Q_r matrix.

This hierarchy presents a cognitive model of task performance for skills in the areas of ratio, factoring, function, and substitution. The hierarchy contains two independent

branches which share a common prerequisite, attribute A1. Aside from attribute A1, the first branch includes two additional attributes, A2 and A3, and the second branch includes a self-contained sub-hierarchy which includes attributes A4 through A9. Three independent branches compose the sub-hierarchy: attributes A4, A5, A6; attributes A4, A7, A8; and attributes A4, A9.

As a prerequisite attribute, attribute A1 includes the most basic arithmetic operation skills, such as addition, subtraction, multiplication, and division of numbers. Attributes A2 and A3 both deal with factors. In attribute A2, the examinee simply needs to have knowledge about the property of factors. In attribute A3, the examinee not only requires knowledge of factoring (i.e., attribute A2), but also the application of factoring. Therefore, attribute A3 is considered a more advanced attribute than A2. The self-contained sub-hierarchy contains six attributes. Among these attributes, attribute A4 is the prerequisite for all other attributes in the sub-hierarchy. Attribute A4 has attribute A1 as a prerequisite because A4 not only represents basic skills in arithmetic operations (i.e., attribute A1), but it also involves the substitution of values into algebraic expressions which is more abstract and, therefore, more difficult than attribute A1. The first branch in the sub-hierarchy deals, mainly, with functional graph reading. For attribute A5, the examinee must be able to map the graph of a familiar function (e.g., a parabola) with its corresponding function. Attribute A6 deals with the abstract properties of functions, such as recognizing the graphical representation of the relationship between independent and dependent variables. The second branch in the sub-hierarchy considers the skills associated with advanced substitution. Attribute A7 requires the examinee to substitute numbers into algebraic expressions. The complexity of attribute A7 relative to attribute A4 lies in the concurrent management of multiple pairs of numbers and \ multiple equations. Attribute A8 also represents the skills of substitution. However, what makes attribute A8 more difficult than attribute A7 is that algebraic expressions, rather than numbers, need to be substituted into another algebraic expression. The last branch in the sub-hierarchy

contains only one additional attribute, A9, related to skills associated with rule substitution. It is the rule, rather than the numeric value or the algebraic expression, that needs to be substituted in the item to reach a solution.

SAT Example 1: Training without Extra Output

In the first example, training was conducted without extra output. That is, the input to train the network is the expected response vectors produced from the AHM feature creation analyses and the output is the specific combination of examinee attributes derived from the transpose of the Q_r matrix for each expected response vector. The relationship between the expected response vectors with their associated attribute vectors was established by presenting each pattern to the network repeatedly.

Using nine hidden units, the network converged using a model with 9 input, 9 hidden, and 9 output units. The value for the root mean square was 0.00082 after 500 epochs. The probabilities associated with each attribute across the nine expected response patterns was used to define the functional relationship for mapping the examinees' observed response pattern onto the expected response pattern so their attribute mastery levels could be determined.

Seven examples are presented in Table 1. The first three include attribute probabilities for observed response patterns that are consistent with the cognitive model in Figure 1. Take, for instance, an examinee who possesses the first three attributes, A1 to A3, thereby producing the response pattern 111000000 (i.e., example 1). This observed response pattern is consistent with one of the 58 expected response patterns. The attribute probabilities for this response pattern are 0.91, 1.00, 1.00, 0.08, 0.02, 0.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively. Examples 2 and 3 illustrate the attribute probabilities associated with observed response patterns that are also consistent with the hierarchy in Figure 1.

Alternatively, examples 4 to 7 illustrate attribute probabilities for observe response patterns that are *inconsistent* with the attribute hierarchy. In other words, these response patterns are not one of the 58 patterns in

Figure 1.

expected response matrix. These inconsistency can be addressed using the network because its purpose is to define the functional relationship for mapping the examinees' observed response pattern onto the expected response pattern using $M_k^* = F(a_k)$.

The first inconsistent pattern, example 4, includes examinees who correctly solve the items associated with attributes A1 and A3, but then incorrectly solve the item associated with attribute A2. According to the cognitive model in Figure 1, this response patterns is not expected because A3 requires A1 and A2. Yet, we have an observed response pattern where A3 is solved correctly while A2 is not. This inconsistency or slip means that the examinee's item response is unexpected because the attributes probed by the item are assumed to be mastered by the examinee, given the cognitive model of task performance. The attribute probabilities for this observed response pattern are 0.92, 0.99, 1.00, 0.16, 0.04, 0.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that it is very unlikely that an examinee who possesses attribute A3 would not also possess attribute A2, if the cognitive model

in Figure 1 is true. The attribute probability level is also unusually high, in this example, because we only have one item measuring each attribute and this branch (A1 to A3) has only three attributes, in total. However, when a larger number of items are used to measure the attributes across a larger number of branches, the attribute probabilities decrease, as illustrated in examples 5 to 7.

For these three examples, attribute A4, which is the prerequisite attribute in each case, is missing. In example 5, the examinee correctly solves the items measuring A1, A5 and A6, but incorrectly solves the item measuring A4. The attribute probabilities for this observed response pattern are 0.69, 0.01, 0.00, 0.31, 1.00, 1.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A5, and A6, but likely not A4. A value of 0.50 is used in our example to interpret the probabilities, meaning that if the probability is greater than 50%, the examinee is believed to possess the attribute. In example 5, however, it is difficult to evaluate A4 because the examinee only solves two items correctly that required A4. In example 6, on the other hand, the examinee

correctly solves the items measuring A1 and A5 to A8. In this case, four items that require A4 are correctly solved. The attribute probabilities for this observed response pattern are 0.43, 0.00, 0.00, 0.95, 1.00, 1.00, 0.99, 0.78, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A4 to A8. The examinees may also possess A1, but the probability is low (the result for A1 in this example is unusual because the examinee must possess A1 to solve the remaining items). Notice that when all four items requiring the prerequisite attribute are correctly solved (i.e., A5 to A8), but the prerequisite attribute is incorrectly solved (i.e., A4), the probability is high that the examinee, in fact, possesses the prerequisite A4. Or, stated differently, it is unlikely that the examinee could solve the items associated with A5 to A8 without possessing A4, if the cognitive model in Figure 1 is accurate. When the final attribute is included, A9, in example 7, the attribute probabilities are 0.87, 0.01, 0.00, 0.96, 1.00, 0.99, 0.97, 0.62, 0.98 indicating that the examinee possesses A1,

A4 to A9. The results across the seven examples are consistent with our expectations based on the cognitive model, for the most part. The only unusual results occurred in example 5 where the probability for A4 was unexpectedly low and example 6 where the probability for A1 was also low.

SAT Example 2: Training with Extra Output

In the second example, training was conducted with extra output (Gällmo & Carlström, 1995). That is, the input to train the network is the expected response vectors produced from the AHM feature creation analyses, as in example 1, but the target output is the specific combination of examinee attributes derived from the transpose of the Q_r matrix as well as the ability estimate for each expected response vector.

With a cognitive diagnostic model like the AHM, expected item and ability parameters can be estimated. The expected item parameters can be produced using an item response theory (IRT) model. For example 2, the two-parameter (2PL) logistic IRT model is used. This model is given by

Table 1. Attribute Probabilities for Seven Observed Examinee Response Patterns using the SAT Algebra Hierarchy in Figure 1 with No Extra Output

<i>Pattern</i>	<i>Attribute Probability</i>								
	1	2	3	4	5	6	7	8	9
<i>Consistent</i>									
1. A1 to A3	0.91	1.00	1.00	0.08	0.02	0.00	0.00	0.00	0.00
2. A1, A4 to A6	0.94	0.01	0.00	0.96	1.00	0.97	0.01	0.00	0.00
3. A1, A4 to A8	0.96	0.00	0.00	1.00	1.00	0.97	1.00	0.98	0.02
<i>Inconsistent</i>									
4. A1, A3 (Missing A2)	0.92	0.99	1.00	0.16	0.04	0.00	0.00	0.00	0.00
5. A1, A5, A6 (Missing A4)	0.69	0.01	0.00	0.31	1.00	1.00	0.00	0.00	0.00
6. A1, A5 to A8 (Missing A4)	0.43	0.00	0.00	0.95	1.00	1.00	0.99	0.78	0.00
7. A1, A5 to A9 (Missing A4)	0.87	0.01	0.00	0.96	1.00	0.99	0.97	0.62	0.98

$$P(u = 1|\Theta) = \frac{1}{1 + e^{-1.7a_i(\Theta - b_i)}}, \quad (7)$$

where a_i is the item discrimination parameter, b_i is the item difficulty parameter, and Θ is the ability parameter. Using the 2PL logistic IRT function, the a and b parameters can be determined for each item using the expected item response patterns given by the columns of the expected response matrix. The expected ability parameters are then produced by locating the maximum of the likelihood function defined by

$$L(\mathbf{u}|\theta_j) = \prod_{i=1}^n P_{ij}^{u_{ij}} Q_{ij}^{1-u_{ij}} \quad (8)$$

where $P_{ij}^{u_{ij}}$ is the probability, based on the 2PL logistic function, for a correct response to item i and $Q_{ij}^{1-u_{ij}}$ is $1 - P_{ij}^{u_{ij}}$. The likelihood function is typically placed on a unidimensional scale with a mean of 0 and a standard deviation of 1.

To illustrate the extra output training method, a random sample of 5000 simulated examinees was generated for the 58 unique patterns in the expected response matrix with the constraint that the distribution of total score be normal in shape. Then, the simulated response data were fit to the 2PL logistic IRT model to estimate the item and ability parameters. Estimation was conducted with the computer software BILOG-MG (du Toit, 2003). Default settings in BILOG-MG were used, except the calibration option that was set to “float” indicating that the means of the priors on the item parameters were calculated using marginal maximum likelihood estimation, and both the means and the item parameters were updated after each iteration. The ability estimates provide a measure of the expected examinees’ score on a (0, 1) unidimensional scale which typically ranges from -4 to +4. Thus, a higher score indicates a higher ability level.

These ability scores have an important role in the example 2 analysis: They serve as extra output or “hints” that provide prior knowledge to the neural network about a feature in each expected response pattern that may

increase the accuracy of learning. The ability level extra output is only included to help the network learn, and once training is complete, the extra output is removed. The benefit of adding an extra output, like ability level, is that it can act as a side constraint thereby increasing the representational power of the network and potentially increase the accuracy and generalizability of the network solution.

Using nine hidden units, the network converged using a model with 9 input, 9 hidden, and 9 output units. The value for the root mean square was 0.00028 after 500 epochs. The probabilities associated with each attribute across the nine expected response patterns was used to define the functional relationship for mapping the examinees’ observed response patterns from the SAT dataset onto the expected response patterns derived from the cognitive model so their attribute mastery levels can be determined. The attribute probabilities for the same seven response patterns in Table 1 are presented in Table 2.

The results between Tables 1 and 2 are similar, except for two important exceptions. Recall, for example 5 in Table 1, the examinee correctly solved the items measuring A1 and A5, but incorrectly solved the item measuring A4. The attribute probabilities for this observed response pattern was 0.69, 0.01, 0.00, 0.31, 1.00, 1.00, 0.00, 0.00, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A5, and A6, but not A4. The same example, but with extra output, shown in Table 2, yields a more interpretable result. The attribute probabilities are 0.95, 0.01, 0.00, 0.60, 1.00, 0.99, 0.01, 0.01, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1 and A5, and likely possesses A4, which is expected given that the examinee correctly solved the item measuring A5. In Table 1, example 6, the attribute probability for A1 was low, given that the examinee required this attribute to solve the items. But, in Table 2, example 6, the attribute probabilities are more consistent with the cognitive model at 0.97, 0.01, 0.00, 0.95, 1.00, 0.99, 1.00, 0.99, and 0.00 for attributes A1 to A9, respectively, indicating that the examinee possesses A1, A4 to A8. When A9 is added in example 7, the attribute probabilities are 0.98, 0.04, 0.00, 1.00, 1.00,

0.98, 1.00, 1.00, 0.98 indicating that the examinee possesses A1, A4 to A9. The probability for A8 in example 1 was reasonably high at 0.62. But, in example 2, the probability for A8 is much higher at 1.00 and, thus, easier to interpret. To summarize, the results across the seven examples in Table 2 are consistent with our expectations based on the cognitive model in Figure 1, particularly when compared to the results in Table 1. These outcomes also reveal that extra output learning improved the interpretability of the network solutions.

Discussion

Assessment engineering with the AHM relies on two stages. In the feature creation stage, principled test design procedures are used to develop items that systematically measure each component in the cognitive model. In the pattern recognition stage, the functional relationship between the examinees' expected response patterns and item attributes is established so the attribute probabilities for the examinees' observed response patterns can be estimated. The purpose of the present study was to describe the analytic procedures in the pattern recognition stage.

Using response data from a sample of examinees who wrote algebra items on the SAT, the results from two different examples were presented. In the first example, the attribute probabilities were computed by training the network without extra output. The value for the root mean square was small at 0.00082. The results across the seven examples were consistent with our expectations from the cognitive model, for the most part, as only two anomalous results were noted. In the second example, the attribute probabilities were computed by training the network with extra output associated with the ability estimates for each expected response pattern. The ability estimates served as an excellent source of extra learning output because they were derived from an IRT model fit to the expected response patterns to produce a single score for each unique pattern. The network yielded a smaller root mean square (0.00028) compared to the network without extra output, and the results across all seven examples were consistent with

the cognitive model indicating that extra output training increased the interpretability of the network solution.

One limitation of the current study stems from the use of a post-hoc or retrofitting approach when identifying and applying the cognitive model of task performance to the algebra items on the SAT. In the current study, we generated a cognitive model of task performance by conducting a content review of the SAT algebra I and II items to identify the mathematical concepts, operations, procedures, and strategies used by students to solve items on the SAT. However, no new items were developed from the cognitive models of task performance used to produce the attribute hierarchies in Figure 3. This decision was made, in part, because the purpose of the study was to describe and illustrate the analytic procedures in the pattern recognition stage. However, in future applications of the AHM, researchers and practitioners implementing the AHM for AE should begin by specifying the cognitive model and use the attribute hierarchy to develop test items. These model-based test items can then be analyzed using the neural network procedures.

In closing, the role that pattern recognition procedures could one-day play in educational measurement is significant. In May 2006, Eduventures, a market research firm that specializes in educational products and applications, claimed that new applications of formative testing, like cognitive diagnostic assessment, may soon emerge to redefine the educational measurement practices in American classrooms. But they also noted that this emergence will only occur when several key objectives are met, including "the building of truly advanced analytic capabilities, relying on a neural network architecture to act as the engine to convert assessment inputs into prescriptive action" (Wiley, 2006). Our study provides one example of the "advanced analytic capabilities" that are possible when psychometric methods like the AHM incorporate pattern recognition procedures to classify examinees' response patterns on educational tests.

TWO DIMENSION MARGINAL DISTRIBUTIONS OF CROSSING TIME

Table 2. Attribute Probabilities for Seven Observed Examinee Response Patterns using the SAT Algebra Hierarchy in Figure 1 with Ability as Extra Output

<i>Pattern</i>	<i>Attribute Probability</i>									<i>Ability</i>
	1	2	3	4	5	6	7	8	9	
<i>Consistent</i>										
8. A1 to A3	0.96	1.00	0.98	0.04	0.00	0.00	0.01	0.01	0.00	-2.408
9. A1, A4 to A6	0.99	0.01	0.00	0.98	1.00	0.98	0.01	0.00	0.01	-0.001
10. A1, A4 to A8	0.98	0.01	0.00	0.99	1.00	0.99	1.00	0.99	0.02	1.205
<i>Inconsistent</i>										
11. A1, A3 (Missing A2)	0.95	0.99	0.95	0.06	0.00	0.00	0.01	0.00	0.00	-2.408
12. A1, A5, A6 (Missing A4)	0.95	0.01	0.00	0.60	1.00	0.99	0.01	0.01	0.00	-0.369
13. A1, A5 to A8 (Missing A4)	0.97	0.01	0.00	0.95	1.00	0.99	1.00	0.99	0.00	1.072
14. A1, A5 to A9 (Missing A4)	0.98	0.04	0.00	1.00	1.00	0.98	1.00	1.00	0.98	1.429

Acknowledgement

The research reported in this study was conducted with funds provided to the first author by the College Board (Contract 2005-013) and by the Social Sciences and Humanities Research Council of Canada (SSHRC). We would like to thank the College Board and SSHRC for their support. The authors are solely responsible for the methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of the College Board or SSHRC.

References

- du Toit, M. (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT* [Computer Programs], Lincolnwood, IL : Scientific Software.
- Gällmo, O, & Carlström, J. (1995). Some experiments using extra output learning to hint multi layer perceptrons. In L. F. Niklasson, M. B. Boden (Eds.), *Current trends in connections—Proceedings of the 1995 Swedish conference on connections* (pp. 179-190). Mahwah, NJ: Erlbaum.
- Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement*, 44, 325-340. Invited Paper appearing in Special Issue of IRT-Based Cognitive Diagnostic Models and Related Methods, Lou DiBello & William Stout (Guest Editors).
- Gierl, M. J., Leighton, J. P., & Hunka, S. (2007). Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills. In J. P. Leighton & M. J. Gierl (Eds.), *Cognitive diagnostic assessment in education: Theory and applications*. (pp. 242-274). Cambridge, UK: Cambridge University Press.
- Gierl, M. J., Leighton, J. P., Wang, C., Zhou, J., Gokert, R., & Tan. A. (2006, December). *Validating cognitive models of task performance in algebra on the SAT®*. New York: College Examination Board.
- Gierl, M. J., Wang, C., & Zhou, J. (2006, June). *Using the attribute hierarchy method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT®*. New York: College Examination Board.

Leighton, J. P., & Gierl, M. J. (Eds.) (2007a). *Cognitive diagnostic assessment for education: Theory and Applications*. Cambridge, UK: Cambridge University Press.

Leighton, J. P., & Gierl, M. J. (2007b). Defining and evaluating models of cognition used in educational measurement to make inferences about examinees' thinking processes. *Educational Measurement: Issues and Practice*, 26, 3-16.

Leighton, J. P., Gierl, M. J., & Hunka, S. (2004). The attribute hierarchy model: An approach for integrating cognitive theory with assessment practice. *Journal of Educational Measurement*, 41, 205-236.

Luecht, R. M. (2006, September). *Assessment engineering: An emerging discipline*. Paper presented in the Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada.

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational Models of Cognition*. Oxford: Oxford University Press. 21-53.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986a). Learning representations by back-propagating errors. *Nature*, 323, 533-536.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986b). *Parallel distributed processing* (Vol. 1). Cambridge, MA: MIT Press.

Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd Edition., pp. 263-331). New York: American Council on Education, Macmillian.

Wiley, T. (2006, May). *Formative instruction and the quest for the killer application*, Eduventures, Boston, MA.

Robustness of Some Estimators of Linear Model with Autocorrelated Error Terms When Stochastic Regressors are Normally Distributed

Kayode Ayinde, Ladoke Akintol
University of Technology

J. O. Olaomi
University of Ibadan

Performances of estimators of the linear model under different level of autocorrelation (ρ) are known to be affected by different specifications of regressors. The robustness of some methods of parameter estimation of linear model to autocorrelation are examined when stochastic regressors are normally distributed. Monte Carlo experiments were conducted at both low and high replications. Comparison and preference of estimator(s) are based on their performances via bias, absolute bias, variance and more importantly the mean squared error of the estimated parameters of the model. Results show that the performances of the estimators improve with increased replication. In estimating all the parameters of the model, the Ordinary Least Square (OLS) estimator is more efficient than any of the Generalized Least Square (GLS) estimators considered when $-0.25 < \rho \leq 0.25$; and the Maximum Likelihood (ML) and the Hildreth and LU (HILU) estimators are robust.

Key words: Robustness, Stochastic regressors, linear model with autocorrelated error, OLS estimator, Feasible GLS Estimators.

Introduction

The Ordinary Least Square (OLS) estimator is unbiased but inefficient in estimating the parameters of the linear model with autocorrelated error terms, and its predicted values are inefficient if the variance of the autocorrelated error terms are underestimated (Johnston, 1984; Fomby et. al, 1984; Maddala, 2002). Consequently, the Generalized Least Square (GLS) estimator was developed.

Aitken (1935) had shown that the GLS estimator given by

$$\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y \quad (1)$$

with the variance – covariance matrix

$$V(\hat{\beta}) = \sigma^2 (X^T \Omega^{-1} X)^{-1} \quad (2)$$

is efficient among the class of linear unbiased estimator provided Ω is known. Consider the linear model where the error terms follow AR (1) process

$$Y = X\beta + U \quad (3)$$

where

$$U_t = \rho U_{t-1} + \varepsilon_t \quad |\rho| < 1 \quad t = 1, 2, \dots, n \quad \varepsilon_t \sim N(0, \sigma^2)$$

Then the inverse of Ω is given as

$$\Omega^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}_{n \times n}$$

Kayode Ayinde, Ladoke Akintola: University of Technology, Department of Pure and Applied Mathematics, P. M. B. 4000, Ogbomoso, Oyo State, Nigeria. Email: bayoayinde@yahoo.com. J.O. Olaomi: University of Ibadan, Department of Statistics, Ibadan, Oyo State, Nigeria. Email: olaomijo@yahoo.com

Cochrane and Orcutt (1949) pointed out that the presence of autocorrelated error terms in Linear Model requires some modifications of the usual least square method of estimation. They suggested a transformation that uses the matrix

$$P = \begin{bmatrix} -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ 0 & 0 & -\rho & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}_{(n-1) \times n}$$

which ignores the first observation of the error terms. Paris & Winstein (1954) showed that the appropriate transformation required for the transformation is

$$Q = \begin{bmatrix} (1-\rho^2)^{\frac{1}{2}} & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1 & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -\rho & 1 \end{bmatrix}_{n \times n}$$

which retains the first observation. The difference in the usage of P and Q can be negligible when n is large, but in small sample investigation such as in this study, the difference may be major. However, they both require ρ to be known before they can be used.

Fomby et al. (1984) and others emphasized that in practice ρ (and hence Ω) is usually unknown but has to be estimated. They indicated that many consistent estimators $\hat{\rho}$ of

ρ (and hence $\hat{\Omega}$ of Ω) can be estimated to have the Feasible Generalized Least Square estimators. Some of the Feasible GLS estimators available in literatures are the Cochrane and Orcutt estimator (1949), Hildreth and Lu estimator (1960), Paris & Winstein estimator (1954), Thornton estimator (1982), Durbin estimator (1960), Theil's estimator (1971), the

Maximum Likelihood estimator and the Maximum Likelihood Grid estimator (Beach and Mackinnon, 1978), some of which use either the P or Q transformation matrix. Furthermore, some have also been incorporated into White's SHAZAM program (White, 1978) and the new version of the time series processor (TSP, 2005).

However, these estimators are known to be asymptotically equivalent but the question on which is to be preferred in small samples is another matter. The question at what value of ρ does the OLS estimator become inefficient when compared with the feasible generalized least square estimators arises, and what transformation is to be preferred are still of concern (Johnson, 1984; Fomby et. al, 1984). Therefore, the finite properties of these estimators are studied through Monte Carlo methods.

Chipman (1979), Kramer (1980), Kleiber (2001) and others observed that the efficiency of these estimators depends on the structure of the regressors that are used. Rao and Griliches (1969) conducted one of the earliest Monte Carlo studies on the performances of some of these estimators with autoregressive stochastic regressor. They observed that the OLS estimator is only more efficient than any of the GLS estimators considered when $|\rho| < 0.3$; and

that the performances of the GLS estimators are not far apart. Park and Mitchell (1980) observed that when regressors are trended, the estimator that uses the P transformation (Paris & Winstein) is more efficient than the one that uses the Q transformation (Cochrane – Orcutt) and that the latter should even be avoided since it is less efficient than the OLS estimator.

More recently, Nwabueze (2005a) examined the performance of some of these estimators with exponential independent variable. His result, among other things, show that the OLS estimator compares favorably with the Maximum Likelihood (ML) and Maximum Likelihood Grid (MLGD) estimators for small value of ρ but it appears to be superior to Cochrane – Orcutt (CORC) and the Hildreth and Lu (HILU) especially when ρ is large. Some other recent works that are done with different specification of regressors include that of

LINEAR MODEL WITH AUTOCORRELATED ERROR TERMS

Iyaniwura and Nwabuwze (2004a), Iyaniwura and Nwabuwze (2004b), Nwabuwze (2005b), Nwabuwze (2005c), and Olaomi and Iyaiwura (2006).

Consequently, without loss of generality, the purpose of this article is to find out if any or some of these estimators would be robust to autocorrelation when stochastic regressors are normally distributed.

Methodology

Consider the GLS model with stochastic regressors and AR (1) of the form

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + u_t \quad (4)$$

where $u_t = \rho u_{t-1} + \varepsilon_t$

$$|\rho| < 1 \quad t=1,2,\dots,n \quad \varepsilon_t \sim N(0, \sigma^2).$$

Its parameter estimations can be done using the OLS and the (feasible) GLS estimators. Thus, the performances of the OLS estimator and the following feasible GLS estimators are studied: CORC, HILU, ML and the MLGD estimators. The CORC and HILU estimators use the P transformation while the ML and MLGD estimator use the Q transformation.

Monte Carlo experiments were performed for $n = 20$, a small sample size representative of many time series study (Park and Mitchell, 1980) with four replication (R) levels (R = 10, 40, 80, 120) and nine various degree of autocorrelation ($\rho = -0.99, -0.75, -0.5, \dots, 0.99$). At a particular choice of ρ and R (a scenario), the first replication was obtained by generating $e_t \sim N(0,1)$ and hence u_t . Assuming the process start from infinite past and continue to operate, the initial value of U (i.e u_1) was thus drawn from a normal population with mean zero and variance $\frac{1}{1-\rho_1^2}$. Hence

$$u_1 = \frac{\varepsilon_1}{\sqrt{1-\rho_1^2}} \quad (6)$$

$$u_t = \rho_1 u_{t-1} + \varepsilon_t \quad t = 2, 3, \dots, 20 \quad (7)$$

Furthermore, $x_{1t} \sim N(0,1)$ and $x_{2t} \sim N(0,1)$ were generated. Hence, the values of y_t in equation (1) were also calculated by setting the true regression coefficients as $\beta_0 = \beta_1 = \beta_2 = 1$. This process continued until all replications in this scenario were obtained. Another scenario then started until all the scenarios were completed.

Evaluation and comparison of estimators were examined using the finite sampling properties of estimators which include bias (B), absolute bias (AB), and variance (Var) and the more importantly the mean squared error (MSE) criteria. For any estimator $\hat{\beta}_i$ of β_i of model (4)

$$\bar{\hat{\beta}}_i = \frac{1}{R} \sum_{j=1}^R \hat{\beta}_{ij} \quad (7)$$

$$B(\hat{\beta}_i) = \frac{1}{R} \sum_{j=1}^R (\hat{\beta}_{ij} - \beta_i) = \bar{\hat{\beta}}_i - \beta_i \quad (8)$$

$$AB(\hat{\beta}_i) = \frac{1}{R} \sum_{j=1}^R |\hat{\beta}_{ij} - \beta_i| \quad (9)$$

$$Var(\hat{\beta}_i) = \frac{1}{R} \sum_{j=1}^R (\hat{\beta}_{ij} - \bar{\hat{\beta}}_i)^2 \quad (10)$$

$$MSE(\hat{\beta}_i) = \frac{1}{R} \sum_{j=1}^R (\hat{\beta}_{ij} - \beta_i)^2 \\ = Var(\hat{\beta}_i) + [B(\hat{\beta}_i)]^2 \quad (11)$$

for $i = 0, 1, 2$ and $j = 1, 2, \dots, R$.

For each of the estimation methods, a computer program was written using TSP software to estimate all the model parameters and to evaluate the criteria. The four replication levels were further grouped into low (R = 10, 40) and high (R = 40, 80) and the effect of autocorrelation on the performances of the methods (estimators) were examined via the Analysis of Variance of the criteria of each of

Table 1. Summary of the ANOVA TABLE showing the sum of squares of the model parameters based on the criteria in the two replication groups.

Parameter	Replication Group	Type III Sum of squares					
		Source	d.f	Bias	Absolute Bias	Variance	Mean Squared Error
β_0	LOW	R	8	22.290***	293.867***	21568.681***	25967.186***
		M	4	.137	7.470E-02	3.560	7.428
		R*M	32	.619	.876	29.330	68.430
		Error	45	22.356	2.308	87.452	591.721
		Total	89	45.403	297.126	21689.023	26634.766
	HIGH	R	8	.319***	258.096***	20470.474***	20505.716***
		M	4	4.858E-03	4.201E-02**	1.596**	1.583*
		R*M	32	1.862E-02	.228***	10.494***	10.380***
		Error	45	2.367E-02	.105	4.651	4.688
		Total	89	.366	258.471	20487.216	20522.367
β_1	LOW	R	8	7.462E-02***	.206***	1.698***	1.763***
		M	4	1.940E-02*	.459**	1.445***	1.539***
		R*M	32	.162***	1.271***	7.632***	8.005***
		Error	45	5.861E-02	2.786E-02	.540	.542
		Total	89	.315	1.964	11.315	11.848
	HIGH	R	8	9.684E-03***	.291***	3.657***	3.678***
		M	4	2.858E-03*	.399***	2.367***	2.393***
		R*M	32	1.917E-02**	1.451***	14.913***	15.036***
		Error	45	9.472E-03	1.861E-03	3.527E-03	5.605E-03
		Total	89	4.118E-02	2.143	20.940	21.113
β_2	LOW	R	8	.264***	.407***	6.920***	8.095***
		M	4	.200***	.550***	4.463***	5.245***
		R*M	32	.310***	1.727***	27.988***	32.496***
		Error	45	5.493E-02	9.957E-02	1.060	.825
		Total	89	.829	2.783	40.431	46.661
	HIGH	R	8	4.806E-02***	.323***	5.885***	6.164***
		M	4	3.751E-02***	.420***	3.580***	3.780***
		R*M	32	7.292E-02**	1.667***	24.255**	25.385***
		Error	45	3.841E-02	2.413E-03	.388	.267
		Total	89	.197	2.412	34.109	35.595

* $\Rightarrow$ Computed F value is significant at $\alpha = 0.05$. ** $\Rightarrow$ Computed F value is significant at $\alpha = 0.01$.

*** $\Rightarrow$ Computed F value is significant at $\alpha = 0.001$. $\rho = R \Rightarrow$ Autocorrelation levels

M $\Rightarrow$ Methods (Estimators)

the model parameters in the two replication groups. Because at least one of the estimators (CORC) is biased in small samples (Rao & Griliches, 1969), and that the mean squared error is known to replace the absolute bias (Kruthkoff, 1970) and also comprises variance and bias; therefore a further test on significant interaction effect of autocorrelation by method was performed on the basis of the mean squared error criterion. The LSD test of the estimated marginal mean was done at each level of autocorrelation.

At a particular level of autocorrelation, estimators were preferred if their estimated marginal means are not significantly different from the most preferred one. An estimator is most preferred if its estimated marginal mean is the smallest. Estimators that are preferred at all the levels of autocorrelation are said to be robust to autocorrelation; and if estimators are robust to autocorrelation in all the model parameters, the estimators are simply said to be robust.

Simulation Results and Discussion

The summary of findings on the performances of the estimators based on the criteria for each of the model parameters in the two replication groups is given in Table 1. It is observed that the error sum of square and hence the mean square error (if estimated) in all the criteria reduce with increased replications. Thus, the performances of the estimators in estimating all the parameters of the model improve with increased replication.

In estimating β_0 , the interaction effect of autocorrelation and method is only significant at the high replication group in all the criteria except bias. Thus, the performances of the methods are not affected by autocorrelation in bias criterion but in others criteria they do. The estimated marginal means based on the mean squared error criterion are shown in appendix. From the appendix, it is observed that as ρ decreases from zero, the estimated marginal means of the GLS estimators decrease while that of the OLS estimator first decreases before it starts to increase. As ρ increases from zero, the estimated marginal means of all the methods increase. Furthermore, the OLS estimator is observed to be more efficient than any of the

GLS estimators when $-0.25 < \rho \leq 0.25$. It is also noted that the ML and the HILU estimators are robust to autocorrelation in estimating β_0 .

In estimating β_1 and β_2 , the interaction effect of autocorrelation and method is significant at the two replication groups in all the criteria. Thus, the performances of the estimators are affected by autocorrelation in all the criteria under the two replication groups. The estimated marginal means based the mean squared error of the estimated parameters under the two replication groups are given in appendix. The estimated marginal mean of the OLS estimator increases as $|\rho|$ increases while that of the GLS estimators decrease as $|\rho|$ increases, although this is not consistently the situation in β_2 especially when replication is low. Furthermore, it is observed that in estimating β_1 the OLS estimator is only more efficient than any of the GLS estimators at the two replication groups when $|\rho| \leq 0.25$ while in estimating β_2 OLS is more efficient when $|\rho| < 0.25$ at the low replication and when $-0.25 < \rho \leq 0.25$ at high replication. Moreover, the GLS estimators are robust in estimating β_1 and β_2 of the Linear Model.

Conclusion

Because performances of the estimators improve with increased replication, it can therefore be concluded that in estimating all the parameters of the model the ML and HILU estimators are robust; and that OLS estimator is more efficient than any of the GLS estimators considered when $-0.25 < \rho \leq 0.25$.

References

- Aitken, A. C. (1935). On least squares and linear combinations of observations. *Proceedings of the Royal Statistical Society. Edinburgh*, 55, 42 – 48.
- Beach, C. M., & Mackinnon, J. S (1978). A Maximum likelihood procedure for regression with autocorrelated errors. *Econometrica*, 46(1), 51 – 57.

- Chipman, J. S. (1979). Efficiency of Least Squares Estimation of Linear Trend when residuals are autocorrelated *Econometrica*, 47, 115 – 127.
- Cochrane, D., & Orcutt, G. H. (1949). Application of least square to relationship containing autocorrelated error terms. *Journal of the American Statistical Association*, 44, 32 – 61.
- Durbin, J. (1960). Estimation of Parameters in time – series regression models. *Journal of the Royal Statistical Society, B*, 22, 139 – 153.
- Fomby, T. B., Hill, R. C., & Johnson, S. R.(1984). Advanced econometric methods. NY: Springer – Verlag.
- Hildreth, C., & Lu, J. Y. (1960). Demand relationships with autocorrelated disturbances. *Agricultural Experiment Statistical Bulletin* 276. East Lansing, MI: Michigan State University.
- Iyaniwura, J.O. & Nwabueze, J.C. (2004a). Estimators of linear model with autocorrelated error terms and trended independent variable. *Journal of Nigerian Statistical Association*, 17: 20 – 28.
- Iyaniwura, J.O. & Nwabueze, J.C. (2004b).Estimating the autocorrelated error model with GNP data. *Journal of Nigerian Statistical Association*, 17: 29 – 32.
- Iyaniwura, & Olaomi (2006). Efficiency of GLS estimators in linear regression model with autocorrelated error terms which are also correlated with regressor. *An International Journal of Biological and Physical Science (Science Focus)*. 11(1), 129 – 133.
- Johnston, J. (1984). Econometric methods (3rd Ed.). NY: McGraw Hill.
- Kleiber, C. (2001). Finite sample efficiency of ols in linear regression models with long – memory disturbances. *Economic Letters*, 72, 131 – 136.
- Kramer, W. (1980). Finite sample efficiency of ordinary least squares in the linear regression model with autocorrelated errors. *Journal of the American Statistical Association*, 81, 150-154.
- Krutchkoff, R. G. (1970). *Probability and Statistical Inference*. London: Gordon and Breach Science Publishers.
- Maddala G. S. (2002). Introduction to econometrics (3rd ed.). NY: Wiley.
- Nwabueze, J.C. (2005a). Performances of estimators of linear auto – correlated model with exponential independent variable. *Journal of Nigerian Association of Mathematical Physics*, 9, 385 – 388.
- Nwabueze, J.C. (2005b). Performances of estimators of linear model with auto – correlated error terms when independent variable is normal. *Journal of Nigerian of Association of Mathematical Physics*, 9, 379 – 384.
- Nwabueze, J.C. (2005c). Performances of estimators of linear models with auto – correlated error terms when the independent variable is autoregressive. *Global Journal of Pure and Applied Sciences*, 11(1), 131- 135.
- Park, R. E., & Mitchell, B. M. (1980). Estimating the autocorrelated error model with trended data. *Journal of Econometrics*, 13, 185 – 201.
- Paris, S. J., & Winstein, C. B. (1954). Trend estimators and serial correlation. Chicago: Unpublished Cowles Commission, Discussion Paper.
- Rao, P., & Griliches, Z. (1969). Small sample properties of several two-stage regression methods in the context of autocorrelation errors. *Journal of the American Statistical Association*, 64, 251 – 272.
- Theil, H. (1971). *Principles of econometrics*. NY: Wiley.
- Thornton, D. L. (1982). The appropriate autocorrelation transformation when the autocorrelation process has a finite past. *Federal Reserve Bank of St. Louis Working Paper*, 82 - 002.
- TSP (2005). *Users Guide and Reference Manual. Time Series Processor*. NY.
- White, K. J. (1978). A general computer program for econometric methods – SHAZAM. *Econometrica*, 46, 239 – 240.

LINEAR MODEL WITH AUTOCORRELATED ERROR TERMS

Appendix

+ $\Rightarrow$ Estimate that is significantly different from the most preferred one at $\alpha = 0.05$

ρ_1	M	Replication = High Estimated Marginal	Replication = Low Estimated Marginal	R replication = High Estimated Marginal	Replication = Low Estimated Marginal	R replication = High Estimated Marginal
		Means: β_0	Means: β_1	Means: β_1	Means: β_2	Means: β_2
-.99	OLS	.467	2.455+	3.343+	4.876+	4.309+
	COCR	1.495E-02	3.846E-02	4.990E-02	3.269E-02	4.547E-02
	HILU	1.474E-02	3.801E-02	4.961E-02	3.195E-02	4.450E-02
	ML	1.491E-02	3.882E-02	5.013E-02	3.217E-02	4.430E-02
	MLGD	1.505E-02	4.066E-02	5.180E-02	3.132E-02	4.318E-02
-.75	OLS	3.889E-02	.156	.154	.180	.170
	COCR	1.961E-02	5.127E-02	5.972E-02	3.624E-02	5.597E-02
	HILU	1.965E-02	5.135E-02	5.991E-02	3.587E-02	5.573E-02
	ML	1.930E-02	4.711E-02	5.899E-02	3.496E-02	5.329E-02
	MLGD	1.931E-02	4.732E-02	5.898E-02	3.446E-02	5.293E-02
-.5	OLS	3.352E-02	8.808E-02	8.571E-02	7.802E-02	9.500E-02
	COCR	2.667E-02	8.085E-02	6.966E-02	3.786E-02	6.443E-02
	HILU	2.664E-02	8.121E-02	6.939E-02	3.831E-02	6.472E-02
	ML	2.555E-02	5.732E-02	6.496E-02	3.717E-02	6.206E-02
	MLGD	2.545E-02	5.614E-02	6.438E-02	3.733E-02	6.205E-02
-.25	OLS	3.828E-02	6.348E-02	6.027E-02	4.305E-02	6.943E-02
	COCR	3.862E-02	.110	7.478E-02	3.613E-02	7.206E-02
	HILU	3.848E-02	.110	7.427E-02	3.643E-02	7.208E-02
	ML	3.605E-02	7.062E-02	6.652E-02	3.866E-02	6.864E-02
	MLGD	3.592E-02	7.074E-02	6.641E-02	3.933E-02	6.864E-02
0	OLS	5.191E-02	5.803E-02	5.130E-02	2.983E-02	5.654E-02
	COCR	5.858E-02	.112	7.607E-02	3.812E-02	7.242E-02
	HILU	5.877E-02	.113	7.621E-02	3.832E-02	7.241E-02
	ML	5.299E-02	7.658E-02	6.765E-02	3.039E-02	6.447E-02
	MLGD	5.306E-02	7.595E-02	6.731E-02	3.002E-02	6.405E-02
.25	OLS	8.332E-02	6.888E-02	5.651E-02	2.732E-02	5.110E-02
	COCR	9.563E-02	.103	6.786E-02	4.053E-02	6.463E-02
	HILU	9.649E-02	.103	6.807E-02	3.995E-02	6.518E-02
	ML	8.512E-02	7.588E-02	6.108E-02	2.549E-02	5.632E-02
	MLGD	8.523E-02	7.998E-02	6.181E-02	2.547E-02	5.688E-02
.5	OLS	.169	.102	8.018E-02	4.270E-02	5.670E-02
	COCR	.204	7.069E-02	4.837E-02	3.784E-02	5.399E-02
	HILU	.214	7.025E-02	4.845E-02	3.799E-02	5.443E-02
	ML	.169	5.501E-02	4.556E-02	2.237E-02	4.720E-02
	MLGD	.168	5.247E-02	4.464E-02	2.206E-02	4.688E-02
.75	OLS	.571	.180	.133	.126	9.479E-02
	COCR	.936	4.010E-02	3.969E-02	3.493E-02	4.611E-02
	HILU	.861	3.989E-02	3.975E-02	3.485E-02	4.577E-02
	ML	.549	2.681E-02	3.800E-02	2.284E-02	4.071E-02
	MLGD	.546	2.550E-02	3.752E-02	2.284E-02	4.036E-02
.99	OLS	48.841+	.320+	.208+	.329+	.202+
	COCR	49.897+	3.654E-02	3.466E-02	4.112E-02	3.995E-02
	HILU	46.888	3.479E-02	3.396E-02	3.954E-02	3.967E-02
	ML	47.537	2.156E-02	3.373E-02	2.539E-02	3.758E-02
	MLGD	47.727+	2.171E-02	3.391E-02	2.542E-02	3.746E-02

Confidence Intervals Based on Robust Estimators

Meral Cetin Serpil Aktas
Hacettepe University

Classical estimation of confidence intervals based on the sample mean and variance is sensitive to outliers. Robust methods were proposed for reducing the influence of outliers. The Minimum Volume Ellipsoid estimator (MVE), having a high breakdown point, is one of the robust estimators for location and scale parameters. The robust confidence interval for location parameter is constructed based on the MVE, and compared with the proposed robust confidence interval estimation methods. The performance of the robust confidence interval based on MVE is illustrated with a simulation study. The lengths of 100(1- α)% confidence intervals were investigated.

Key words: Robust estimators, minimum volume ellipsoid estimator, robust confidence interval.

Introduction

Little attention has been given to confidence intervals (CI) based on robust estimators when the underlying distribution is nonsymmetric. Robust confidence limits were studied by Huber (1968). Du Mond and Lenth (1987) studied the robust confidence interval for the biweight M estimator. Tiku and Tan (1986) used the MML estimators for location and compare it with the trimmed mean, Sprout's estimators and the others. Fraiman et.al (2001) constructed the robust confidence interval based on optimal robust M-estimates of location.

Robust confidence interval for the median were given by Staudte and Sheather (1990). Adrover and Zamar (2000) defined the globally robust confidence interval and p-values for the location. The robust confidence interval based on the minimum volume ellipsoid estimator is investigated for location parameter. It is compared with the confidence interval

based on sample mean, Huber estimator and median.

Robust Confidence Interval

The most common technique for finding a 100(1- α)% confidence interval is

$$\bar{x} \pm t_{1-\alpha/2; n-1} [S/\sqrt{n}] \quad (1)$$

where t is the 100(1- α /2) percentage point of the distribution on $n-1$ degrees of freedom. If the data come from the nonsymmetric distribution the confidence interval may vary one sample to another. The sample mean and variance would be affected from the distribution and tend to give long intervals.

Robust confidence interval for median is given by the following

$$\text{Median} \pm t_{1-\alpha/2; n-1} [S^*/\sqrt{n}] \quad (2)$$

where, S^* is the standard error of the median given by (Fraiman et.al, 2001).

$$S^* = [x_a - x_b] / 3.4641 \quad (3)$$

$$a = \left(\frac{n}{2} + \frac{\sqrt{3n}}{2} \right) \text{ and } b = \left(\frac{n}{2} - \frac{\sqrt{3n}}{2} \right)$$

The confidence interval based on Huber estimator is given by

Meral Çetin, Department of Statistics. Her interests are in regression analysis, robust regression and robust model selection. Email: meral@hacettepe.edu.tr. Serpil Aktas, Department of Statistics. Her interests are in categorical data analysis, design of experiment and repeated measurement data analysis. e-mail: spxl@hacettepe.edu.tr.

$$Huber \pm t_{1-\alpha/2; n-1}[S_{huber}] \quad (4)$$

where, Huber is M estimator of location and S_{huber} is the standard error of the huber estimators given by Equation [5],

$$S_{huber} = \frac{mad(x)}{1.486} \quad (5)$$

The MVE, having high breakdown point, is one of the robust estimators for the location and scale parameters (Wilcox, 1997). It is one of several multivariate location and scale estimators. This estimator has high finite-sample breakdown point. The use of estimators with high finite-sample breakdown point yields good performance according to the masking effect.

Rousseeuw (1984) introduced the affine equivariant estimator with maximal breakdown point, by putting $T(X)$ = center of minimal volume ellipsoid covering h points of X where h can be taken equal to $[n+1]/2+1$. This is called the minimum volume ellipsoid estimator (1987). The covariance estimator of this is given by the ellipsoid. Because of the transform $x \rightarrow xA+b$ is an ellipsoid where A and b are the constants, MVE is an affine equivariant estimator, such that any transformation on x does not affect the MVE.

The minimum volume ellipsoid estimator proposed by Rousseeuw (1985) is a robust estimation of location and scale of multivariate data in the presence of outliers. The MVE is the robust estimation of multivariate location and scale defined by minimizing the volume of an ellipsoid containing h points. These robust location and scale estimators can be used to detect multivariate outliers and leverage points.

The MVE estimator searches for the smallest ellipse containing half of the data (Wilcox, 1997). When sampling from a multivariate normal distribution, then it rescaled these estimates so that they estimate the usual population mean and covariance matrix. It is difficult to find the smallest ellipse containing half of data. From the n points, MVE estimator randomly selects h points without replacement and computes the volume of this ellipse. The set of points giving the smallest volume is taken to

be minimum volume ellipsoid. The location and scale MVE estimators yield an effective method for identifying outliers in multivariate data (Rousseeuw, 1990).

This estimator is defined to be the ellipsoid of minimum volume covering at least h points of the data set (Rousseeuw, 1987). The breakdown point of MVE estimator at any p -dimensional sample X is

$$\mathcal{E}_n^*(T, X) = (n/2 - p + 1)/n \quad (6)$$

which converges to 50% as $n \rightarrow \infty$ ((Rousseeuw, 1987). The robust confidence interval for location based on MVE is constructed by,

$$MVE \pm t_{1-\alpha/2; n-1}[S_{mve}] \quad (7)$$

where MVE is the location parameter and the S_{MVE} scale parameter. S_{MVE} is computed as `cov.mve` in the statistical software S-Plus. When the outliers are much larger than the true values S_{MVE} has the best estimation. S_{MVE} estimator takes into account half of the observations which are distributed nearest to an estimated center (Ma & Genton, 2001).

Simulation Study

The performance of the robust confidence intervals of a location parameter is illustrated by Monte Carlo Simulation using the S-Plus coding. Four types of confidence intervals including sample mean, median, Huber and MVE were calculated. Random samples were generated from the normal distribution for the sample sizes $n=25, 50, 100, 500$ and 1000 with 1000 replications. In order to see the effect of outliers on the estimators and also on the confidence interval, the simulation was implemented by generating no outlier, one outlier and many outliers. These outliers were generated for the same samples.

Table 1: Lower - upper Bound and Width of the Confidence Interval of Estimators
When Data Consist of No Outlier

Sample Size	Sample Mean	Huber	MVE	Median
10	5.5737 ; 14.4436 (8.8699)	7.2332; 12.7427 (5.5096)	5.3136;14.6654 (9.3491)	7.8332;12.1622 (4.3290)
25	5.9333;14.0742 (8.1409)	7.3331;12.6631 (5.3300)	4.4203;15.5563 (11.1360)	8.9133;11.1037 (2.1904)
50	6.0959;13.9145 (7.8186)	7.3960;12.6182 (5.2222)	3.8225;16.2043 (12.3817)	9.3179;10.7087 (1.3908)
100	6.0061;13.9147 (7.8287)	7.3842;12.6135 (5.2292)	3.4443;16.5351 (13.0908)	9.5066;10.4798 (0.9732)
500	6.0855;13.9183 (7.8327)	7.3680;12.6357 (5.2647)	3.0507;16.9469 (13.8962)	9.7976;10.1973 (0.3997)
1000	6.0367; 13.9184 (7.8413)	7.3616;12.6335 (5.2719)	2.9606;17.0353 (14.0747)	9.8433;10.1486 (0.3053)

Figure 1: Estimators with respect to the sample size and the length (no outlier).
Length of CI is on the Y axis.

CONFIDENCE INTERVALS BASED ON ROBUST ESTIMATORS

Table 2: Lower - upper Bound and Width of the Confidence Interval of Estimators
When Data Consist of One Outlier

Sample Size	Sample Mean	Huber	MVE	Median
10	-8.875;34.8801 (43.7551)	7.0711;13.6131 (6.5421)	4.0045;15.9879 (11.9834)	7.8139;12.7297 (4.9157)
25	-23.7889;50.9608 74.7498	7.3196;12.9334 5.6139	4.179;15.7979 11.6189	8.8008;11.3764 2.5756
50	-39.0150;66.6043 105.6196	7.3993;12.7218 5.3226	3.6877;16.3001 12.6125	9.3498;10.7582 1.4084
100	-62.6400;90.4405 153.0801	7.3722;12.6985 5.3263	3.3542;16.6501 13.2959	9.5368;10.5112 0.9744
500	-160.5000;188.4520 348.9508	7.3768;12.6395 5.2727	3.0403;16.9475 13.9073	9.7827;10.2144 0.4316
1000	-233.3500;261.3210 494.668	7.3635;12.6357 5.2723	2.9649;7.0233 14.0584	9.8435;10.1498 0.3063

Figure 2: Estimators with respect to the sample size and the lenght (one outlier).
Length of CI is on the Y axis.

Table 3: Lower - upper Bound and Width of the Confidence Interval of Estimators
When Data Consist of Many Outlier

Sample Size	Sample Mean	Huber	MVE	Median
10	-12.8933;44.8948 57.7881	6.9204;14.7930 7.8727	2.7041;17.2674 14.5634	7.5889;13.6491 6.0603
25	-40.9061; 82.5339 123.4400	7.3259;13.6704 6.3445	3.3345;16.6874 13.3529	7.0727;11.6354 2.5628
50	-83.9356;141.9135 225.8391	7.4419;13.3682 5.9263	3.2255;16.7500 13.5245	9.4939;11.0521 1.5582
100	-181.502;279.5056 461.0074	7.3904;13.4107 6.0203	2.9702;17.0183 14.0481	9.7299;10.8072 1.0773
500	-741.4450;960.4278 1701.8720	7.3795;12.9939 5.6441	2.7511;17.2263 14.4752	9.8931;10.3499 0.4568
1000	-1494.2700;1912.8690 3407.3600	7.3829;12.9910 5.6081	2.7648;17.2297 14.4919	9.9703;10.2927 0.3224

Figure 3: Estimators with respect to the sample size and the lenght (many outlier).
Length of CI is on the Y axis.

Robust estimators of location and scale for generated samples were calculated. Classical and robust confidence intervals and the width of the confidence intervals were obtained for 95%. Results were given for the 1000 replications. Lower and upper bounds and the width of the CI are shown in Tables 1-2 and 3 with respect to the sample sizes and the outliers. Figures 1, 2, and 3 also indicate the estimators according to the sample size and outlier(s).

Conclusion

The numerical results given in Tables 1, 2 and 3 show the behaviour of the confidence intervals calculated under different conditions. It is known that the mean is very sensitive to outliers. The width of the confidence intervals for each estimators are not affected by the sample size when data consist of no outlier. While the shortest width is obtained for the median among the others, the widest CI is the MVE. When the sample size is increasing, the width of the median is seen to be decreasing. Unlike others, the width of the CI based on Huber estimates does not vary with the sample size.

The CI based on the mean gives very long width when data consist of one outlier. Other CI's based on robust estimators give similar results for the case of no outlier. Although the CI based on mean yield a large width, the case of more than one outlier, robust CI are not affected by the outliers. Irrespective of the number of outlier, the robust CI give the alike results. Note that the confidence interval based on the MVE estimator is approximately two times wider than the Huber.

It can be concluded that the width of the confidence intervals based on the MVE, Huber and the median are not affected by the outlier(s). In the Figures, note there is no difference between the CI's when the data consist of one outlier and more than one outlier. For large samples the confidence intervals for the Huber and MVE is stationary for the case of outlier. When the distribution is nonsymmetric, utilization of the robust confidence intervals would be appropriate. The smallest CI is always obtained for the case. It should be noted that explicit inferences were not made.

References

- Adrover, J. G., & Zamar, R. H., (2000). Bias robustness of three median-based regression estimates. *Technical Report* No. 194, Department of Statistics, University of British Columbia, Canada.
- Du Mond, & Lenth, R. V. (1987). A Robust Confidence Interval for Location, *Technometrics*, 29(2), 211-219.
- Fraiman, R., Yohai, V. J., Zamar, R., (2001), Optimal robust M-estimates of location, *Annals of Statistics*, 29(1), 194-223.
- Huber, P. J., (1968), Robust confidence limits. *Z.Wahrsch.verw, Geb. 10*, 269-278.
- Ma, Y., & Genton, G. M. ,(2001), Highly robust estimation of dispersion matrices, *Journal of Multivariate Analysis*, 78, 11-36.
- Rousseeuw, P. J., (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P. J, (1985), Multivariate estimation with high breakdown point. *Mathematical Statistics and Applications*, Vol B. W. Grossmann, G. Pflug, I. Vincze, and W. Wertz, Dordrecht, (Eds.): Reidel, 283-297.
- Rousseeuw, P. J., Leroy, A. M., (1987). *Robust regression and outliers detection*. NY: Wiley.
- Rousseeuw, P. J, Van Zomeren, B.C., (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85, 633-639.
- Sachs, L., (1982). *Applied statistics: A handbook of techniques*. Springer-Verlag, New York.
- Staudte, R., G., Sheather, S. J., (1990), *Robust estimation and testing*. NY: Wiley.
- Tiku, M. L, Tan W. Y., & Balakrishnan N. (1986)., *Robust inference*. NY: Marcel.
- Wilcox, R., R., (1997), *Introduction to robust estimation and hypothesis testing*, Academic Pres, San Diego.

Test for Spatio-Temporal Counts Being Poisson

Haiyan Chen
University of Maryland

Howard H. Stratton
SUNY

The new Log-Linear Test (T_L) is proposed to identify when the Poisson model fails for a collection of count random variables. T_L is shown to have better rejection rate with small sample size and essentially the same power compared to a classical Fisher-Bohning's Statistic T_F for standard alternatives to Poisson.

Key words: Fisher-Bohning's Statistic, log-linear test, over-dispersion to Poisson, Negative binomial, Zero-inflated Poisson.

Introduction

Human disease data are often in the form of count data and its associated rate. Examples include disease incidence, prevalence, and/or mortality (Lindsey, 1995, Hinde & Demetrio, 1998). The Poisson distribution is a traditional probability model for count data (Hinde & Demetrio, 1998), and has the property that its expected value equals its variance, i.e., $E(Y) = \text{var}(Y)$. Thus, count data for which $E(Y) < \text{var}(Y)$ indicate over-dispersion relative to the Poisson. The extra disparity could be due to heterogeneity in the population, or an overabundance of certain specific values, e.g., excess zeros (Tiago de Oliveria, 1965, Bohning, 1994, Lambert & Roeder, 1995, Lindsey, 1995, Hinde & Demetrio, 1998, Brown & Zhao, 2002, Smyth,

2003). The term 'over-dispersion' is reserved for over-dispersion relative to the Poisson, i.e., any random variable that has variance-mean ratio greater than one is called over-dispersed. Failure to take account of this over-dispersion can lead to serious underestimation of standard errors and misleading inference for the regression parameters.

Suppose that I independent count variables Y_i ($i = 1, \dots, I$) are each observed N times. The associated I sample means are given by $\bar{Y}_i = \sum_j Y_{ij} / N$ and I sample variances are given by $S_i^2 = \sum_j (Y_{ij} - \bar{Y}_i)^2 / (N - 1)$, where $j = 1, \dots, N$. Hypothesis tests that an individual count variable is Poisson have been developed (Hoel, 1943, Tiago de Oliveria, 1965, Cameron & Trivedi, 1990, Bohning, 1994, Lambert & Roeder, 1995, Lindsey, 1995, Hinde & Demetrio, 1998, Brown & Zhao, 2002, Smyth, 2003), however, little development of tests of hypothesis that a group of count variables are simultaneously distributed as Poisson has been done. This paper investigates three possible hypothesis tests that a group of variables are simultaneously Poisson vs. over-dispersion to the Poisson. It will be seen that only one of these tests is feasible in terms of both test size and test power for relatively small number of observations for each of the variables.

The main data set used to illustrate the proposed tests in this paper, named *NYSLD*, was derived from the New York State Department of Health Lyme Registry Surveillance System. Only confirmed cases using the Lyme disease

Haiyan Chen is a Research Assistant Professor in the Division of Health Services Research in the Department of Health Promotion and Policy in the Dental School. His research interests include Bayesian methods applied to clinical trial design, spatio-temporal data analysis, and survey design and data analysis. Email: hchen@umaryland.edu. Howard Stratton is a Professor in the Department of Epidemiology and Biostatistics in the School of Public Health, SUNY. His research interests include measurement error models, selection bias correction, spatial statistics, mixture models, and hidden Markov chains. Email: hstratton2001@yahoo.com

(LD) surveillance definition (White, Chang, Benach, et al., 1991, CDC, 1997) were selected. For each case, county of residence and year of onset were used. Cases with either of these two pieces of information missing were excluded. The LD data of three neighboring states of New York were accessed online from the Connecticut Department of Public Health (Connecticut State Department of Health website, 2004), New Jersey Department of Health and Senior Services (New Jersey Department of Health and Senior Services website, 2004), and Pennsylvania Department of Health (Pennsylvania Department of Health website, 2004) websites.

A standard reference statistic: the Fisher-Bohning Statistic

A simple diagnostic test for over-dispersion of a single variable has been a long sought goal for deciding whether a further investigation of latent heterogeneity is necessary. Tiago de Oliveira (1965) approached this via the difference of $D_i = (S_i^2 - \bar{Y}_i)$ for an individual random variable Y_i . They argued that D_i ' variance under the null hypothesis $Y_{ij} \sim \text{Poisson}(\lambda_i)$ is given by $(1 - 2\lambda_i^{1/2} + 3\lambda_i)/N$, which can be estimated by $(1 - 2\bar{Y}_i^{1/2} + 3\bar{Y}_i)/N$. The proposed test statistic,

$O_T = N^{1/2} (S_i^2 - \bar{Y}_i) / (1 - 2\bar{Y}_i^{1/2} + 3\bar{Y}_i)^{1/2}$, was treated as if it had a standard Normal limiting null distribution.

However, Dankmar Bohning (1994) showed by simulation that the limiting distribution of Tiago de Oliveira's statistic under the Poisson null hypothesis is neither a standard normal nor is it independent of λ_i . Bohning noted that the failure of Tiago de Oliveira's test is due to incorrect computation of the standard deviation of D_i and showed its correct variance to be $2\lambda_i^2 / (N-1)$. The corrected test statistic,

$$O_T^n = \frac{S_i^2 - \bar{Y}_i}{\sqrt{2\bar{Y}_i^2 / (N-1)}} = \{(N-1)/2\}^{1/2} \{(S_i^2 / \bar{Y}_i) - 1\}$$

is asymptotically $N(0, 1)$.

To address the multiple comparison problem in this paper, the overall p -value for the

I independent Bohning's over-dispersion tests is calculated using the Fisher's statistic for combining independent tests (Hedges & Olkin, 1995) and is named the Fisher-Bohning's Statistic (T_F).

If p_i is p -value of i th individual test for a continuous test statistic, p_i has a uniform $(0, 1)$ distribution when the null hypothesis H_{0i} of the test is true. Fisher's procedure then uses the fact that $-2\log p_i$ has a χ^2 distribution with two degrees of freedom. Because the sum of independent χ^2 variables has a χ^2 distribution with degree of freedom equal to sum of the degrees of freedom of each individual χ^2 , the Fisher-Bohning's Statistic,

$$T_F = -2\log(p_1 p_2 \cdots p_n) = -2\sum_{i=1}^I \log p_i, \text{ has}$$

a χ^2 distribution with $2n$ degrees of freedom under the null hypothesis H_0 . Although the null hypothesis H_{0i} : $Y_i \sim \text{Poisson}$ involves a discrete distribution, it is well approximated by continuous normal distributions if the expected values of the corresponding Poissons are sufficiently large (Johnson, Kotz, & Kemp, 1992). Thus the χ^2 null distribution for Fisher-Bohning's statistic T_F is applicable to the Poisson null hypothesis in this case.

Two new test statistics

Before presenting new test procedures, there are some general concepts and theorems that need to be introduced. First, the concepts 'corresponding zero-inflated variable' and 'corresponding zero-inflated distribution' are defined.

Let Y be a random variable with probability function $p(Y)$ and ω be a value between 0 and 1. If a random variable $\tilde{Y}$ has

$$\tilde{Y} = \begin{cases} 0 & \text{with probability } \omega \\ Y & \text{with probability } 1 - \omega \end{cases}$$

then $\tilde{Y}$ has density function:

$$p(\tilde{Y}) = \begin{cases} \omega + (1-\omega)p(0) & \tilde{Y} = 0 \\ (1-\omega)p(y) & \tilde{Y} = y > 0 \end{cases}$$

and is called the corresponding zero-inflated variable to Y with zero-inflated distribution, $p(\tilde{Y})$.

Theorem 1: If $E(Y^k)$ exists, then $E(\tilde{Y}^k) = (1-\omega)E(Y^k)$, and $\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = \frac{\text{var}(Y)}{E(Y)} + \frac{\omega}{1-\omega}E(\tilde{Y})$ for $E(Y) > 0$, $E(\tilde{Y}) > 0$.

Corollary 1: If $\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y)$ holds, then

$$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega + \eta}{1-\omega}E(\tilde{Y}).$$

The theorem and its corollary presented above provide a basis on which new tests of whether a distribution with the property $\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y)$ is over-dispersed to the

Poisson, i.e. $\eta > 0$ are developed. More precisely the two new proposed tests of hypothesis deal with:

1. $\{Y_i, i = 1, \dots, I\}$ which are I independent random variables.
2. For each Y_i , N independent records were observed.
3. Test Y_i being simultaneously Poisson (λ_i) by the null hypothesis

$$H_0: \frac{\text{var}(Y_i)}{E(Y_i)} = 1 \text{ for } E(Y_i) > 0$$

for all i

versus over-dispersion to the Poisson

$$H_1: \frac{\text{var}(Y_i)}{E(Y_i)} = 1 + \eta E(Y_i) \text{ for}$$

$$E(Y_i) > 0, \text{ and } \eta > 0.$$

That is, a test of $\eta = 0$ vs. $\eta > 0$.

A test based on a linear regression of sample variance-mean ratio on the sample mean

Under H_1 ,

$$\frac{E(y_{ij} - E(Y_i))^2}{E(Y_i)} - 1 = \eta E(Y_i). \text{ If } E(Y_i) \text{ is}$$

known, a test for over-dispersion is a t -test for η in the least-squares (LS) regression

$$\frac{S_{Y_i}^2}{E(Y_i)} - 1 = \eta E(Y_i) + \varepsilon_i,$$

where the error term is defined by

$$\varepsilon_i = \frac{S_{Y_i}^2}{E(Y_i)} - \frac{\text{var}(Y_i)}{E(Y_i)}.$$

Since $E(Y_i)$ is unknown,

it is estimated by $\bar{Y}_i$. A linear regression of the

sample variance-mean ratio on sample mean $\frac{S^2}{\bar{Y}} = \beta_0 + \beta_1 \bar{Y} + \varepsilon$ is made so that a test for

over-dispersion or under-dispersion is a test of whether $\beta_l = 0$ by treating the test statistic,

$$T_L = \frac{\hat{\beta}_1 - 1}{\sqrt{\hat{\text{var}}(\beta_1)}}, \text{ as } N(0, 1) \text{ under } H_0. \text{ WLS is}$$

used to estimate regression coefficients and t -test is used to draw statistical inferences. Note that above justification has been intuitively developed rather than by strict logic. It will be shown to be unreliable.

A test based on a linear regression of the log-sample variance on the log-sample mean

An alternative to T_S is suggested by re-expressing the alternative H_1 ,

$$\frac{\text{var}(Y)}{E(Y)} = 1 + \eta E(Y) \text{ for } E(Y_i) > 0, \text{ via a}$$

logarithmic transformation to give

$\log(\text{var}(Y)) = \log(1 + \eta E(Y)) + \log(E(Y))$. The Poisson condition of $\eta = 0$ is then equivalent to $\log(\text{var}(Y)) = \log(E(Y))$.

The unknown $\text{var}(Y)$ and $E(Y)$ are estimated by S^2 and $\bar{Y}$. In order to test

under/over-dispersion to Poisson, a least square fit of $\log S^2 = \beta_0 + \beta_1 \log \bar{Y} + \varepsilon$ is made.

Rejecting either $\beta_0 = 0$ or $\beta_l = 1$ is sufficient to reject the Poisson. Here a test of whether $\beta_l = 1$ is proposed by treating the test statistic,

$$T_L = \frac{\hat{\beta}_1 - 1}{\sqrt{\hat{\text{var}}(\beta_1)}}, \text{ as } N(0, 1) \text{ in rejecting } H_0. \text{ The}$$

TEST FOR SPATIO-TEMPORAL COUNTS BEING POISSON

validity using this distribution under H_0 will be confirmed by simulation studies.

Simulation

General design for the simulations

Simulations were conducted to examine and compare the size and power characteristic of the three proposed tests T_S , T_L , and T_F . Two alternative hypotheses to a null hypothesis of Poisson that will be tested are 1) $H_1: Y_i \sim$ Negative Binomial, i.e. NB (μ_i, ν), and 2) $H_1: Y_i \sim$ Zero Inflated Poisson, ZIP (μ_i, ω).

Four sample sizes, $N = 5, 11, 50$, or 100 are used to resemble data of small, moderate, and large sample size. Each simulation experiment is based on 500 replications. Two nominal α levels, $\alpha = 0.01$, or $\alpha = 0.05$ are evaluated.

In all cases, μ_i is set equal to i th NYS county's observed average annual incidence rates (per 100,000 population) of LD, where $i = 1, \dots, 57$, in order to have a practical sense of how the tests perform relative to the NYSLD data. Note that during the 11-year studied time period, none of the 57 counties in NYS had a zero average annual incidence rate of LD, indicating that every of those counties had at least one case reported in some year. Figure 1a-b graph the empirical distributions of μ_i and $\log(\mu_i)$, respectively. Range of μ_i is from 0.55 to 349.00 and this covers a relative wide range. The distribution of $\log(\mu_i)$ (skewness = 1.40) is much less positively skewed than that of μ_i (skewness = 3.86).

Analysis of test size

Four sets of data from $H_0: Y_i \sim$ Poisson (μ_i) were generated corresponding to four sample sizes, $N = 5, 11, 50, 100$ for each i ($i = 1, \dots, 57$, and see Section 'General design for the simulations' for the values assigned to μ_i). Percentages of rejections of $H_0: Y_i \sim$ Poisson (μ_i) were calculated for the two α s: $\alpha = 0.01$ and $\alpha = 0.05$ in order to evaluate whether sizes of tests were sufficiently close to their nominal α s. The results are summarized in Table 1.

The size of the test T_S for all four sample sizes turns out to be considerably greater than both the nominal sizes of 1% and 5% (Table 1). The size of the test T_L for all four samples sizes

is statistically indistinguishable from both the nominal sizes of 1% and 5%.

The match between the actual and nominal sizes for T_F is different for different sample sizes: for small sample size as $N = 5$, the size of the test T_F turns out to be smaller than both the nominal sizes of 1% and 5%; while for moderate and large sample size ($N = 10, 50$, or 100), the size of the test T_F is statistically indistinguishable from both the nominal sizes of 1% and 5%.

In summary, the match between the actual and nominal size is worst for T_S and best for T_L . When sample size is adequately large, T_F as well as T_L have consistent test sizes.

Power analysis

Because T_S does not have consistent test sizes but T_L and T_F , essentially do, in the following Sections, power investigations are only made for T_L and T_F as a function of increasing sample size or the degree of over-dispersion.

Power analysis under alternative hypothesis $H_1: Y_i \sim$ NB

Under the alternative hypothesis $H_1: Y_i \sim$ NB (μ_i, ν), the probability density is

$$p(Y_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(\nu)\Gamma(y_i + 1)} \left(\frac{\nu}{\nu + \mu_i} \right)^\nu \left(\frac{\mu_i}{\nu + \mu_i} \right) \text{ and}$$

$$\text{its variance-mean ratio is } \frac{\text{var}(Y)}{E(Y)} = 1 + \frac{E(Y)}{\nu}.$$

Taking the logarithm, this equation becomes:

$$\begin{aligned} \log(\text{var}(Y)) &= \log(E(Y)) + \log(1 + E(Y)/\nu) \\ &= g(E(Y)). \end{aligned}$$

In this experiment, the test power is set up as a function of ν , the dispersion parameter, with 25 different values for ν set discretely from 1 to 5000. This simulates the degree of over-dispersion from large ($1 \leq \nu < 50$), moderate ($50 \leq \nu < 500$) to small ($500 \leq \nu < 5,000$) correspondently. At each value of ν , the experiment described is performed.

Figure 1.a Empirical distribution of μ_i s Figure 1b. Empirical distribution of $\log(\mu_i)$ s

The empirical power curves of the two tests (T_L , T_F) as functions of ν for two nominal test sizes ($\alpha = 1\%$, $\alpha = 5\%$) are presented in Figure 2a-d for sample sizes $N = 5, 11, 50$, and 100. Figure 2a-d indicate that the powers of the two tests are nonlinear monotonically decreasing functions of ν , which represents the degree of over-dispersion. When the degree of over-dispersion to Poisson is large (i.e., ν is 50 or less), both tests T_L and T_F have fairly high powers for all four samples sizes and for two nominal test sizes, ranging from 65.4% to 100%. The power decreases dramatically when the degree of over-dispersion decreases (i.e., ν increases from 50 to 1000). When ν is as big as 5,000, the test powers are very low, ranging from 0.8 to 21.6 (Fig. 2a-d). Sample size seems to have less influence on powers of the tests than the degree of over-dispersion does. When sample size is increased from small ($N = 5$), moderate ($N = 11, 50$) to large ($N = 100$), the corresponding test powers only slightly increase.

With small sample sizes, the ratios between the power of T_L and the power of T_F with the increase of values of ν are relative unstable (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 2a-b). With larger sample size, the power of T_L decreases fast then the power of T_F with the increase of values of ν (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 2c-d), indicating T_L is more sensitive than T_F to the degree of over-dispersion. This is especially true for moderate degree of over-dispersion.

Power analysis under alternative hypothesis H_1 : $Y_i \sim \text{ZIP}$

Under the alternative hypothesis H_1 : $Y_i \sim \text{ZIP}(\mu_i, \omega)$, the probability density is

$$\Pr(Y = y) = \begin{cases} \omega + (1 - \omega) \exp(-\mu) \\ (1 - \omega) \exp(-\mu) \mu^y / y! \end{cases}.$$

The variance-mean ratio is $\frac{\text{var}(Y)}{E(Y)} = 1 + \frac{\omega E(Y)}{1 - \omega}$. Taking logarithms gives,

$$\begin{aligned} \log(\text{var}(Y)) &= \\ \log(E(Y)) + \log(1 + \omega E(Y) / (1 - \omega)) &= \\ g(E(Y)) \end{aligned}$$

In this simulation, the test power is again set up as a function of ω , the dispersion parameter. The empirical power curves of the two tests (T_L , T_F) as functions of ω for two nominal test sizes ($\alpha = 1\%$, $\alpha = 5\%$) are presented in Figure 3a-d for sample sizes $N = 5, 11, 50, 100$ respectively. Total 99 different values, $\omega = \{0.01, 0.02, \dots, 0.98, 0.99\}$, was used while only test powers for $\omega = \{0.01, 0.02, \dots, 0.49, 0.50\}$ are presented in Figure 3a-d. At each value of ω , the experiment described in ‘General design for the simulation’ was performed.

Again, it appears that the test power is a nonlinear monotone increasing function of the degree of over-dispersion to Poisson, which is represented here by ω (Smaller values of ω index smaller degree of over-dispersion). And with small sample size, the ratios of powers of the two tests (T_L , T_F) are unstable with the decrease of values of ω (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 3a-b). With larger sample size, the power of T_L decreases fast then the power of T_F with the increase of values of ν (TL1 vs. TF1 and TL5 vs. TF5 in Fig. 3c-d), indicating T_L is more sensitive than T_F to the degree of over-dispersion.

In summary, the simulation experiments demonstrate that among the three evaluated tests, T_S is ruled out due to unacceptable test size; the power characteristic of T_L is empirically superior to T_F in terms of sensitivity to degree of over-dispersion. Thus, only T_L is used in the four states (New York, Connecticut, New Jersey, and Pennsylvania) LD data including the *NYSLD* data.

Table 1. Percentage rejections of $H_0: Y_i \sim \text{Poisson}(\mu_i)$

For nominal alpha = 1%				
sample size	percentage of reject H0 (95%CI)			
	T_S	T_L	T_F	
5	12.4 (11.6, 13.2)	1.2 (0.4, 2.0)	0.2 (0.0, 1.0)	
11	12.0 (11.2, 12.8)	1.4 (0.0, 1.4)	0.8 (0.0, 1.6)	
50	13.2 (12.4, 14.0)	0.6 (0.0, 1.4)	0.6 (0.0, 1.4)	
100	13.6 (12.8, 14.4)	0.6 (0.0, 1.4)	1.0 (0.2, 1.8)	
For nominal alpha = 5%				
sample size	percentage of reject H0 (95%CI)			
	T_S	T_L	T_F	
5	18.4 (16.4, 20.4)	6.6 (4.6, 8.6)	2.6 (0.6, 4.6)	
11	20.0 (18.0, 22.0)	6.0 (2.4, 6.4)	5.0 (3.0, 7.0)	
50	20.0 (18.0, 22.0)	4.4 (2.4, 6.4)	5.2 (3.2, 7.2)	
100	22.2 (20.2, 24.2)	4.4 (2.4, 6.4)	5.0 (3.0, 7.0)	

Applications

The new test statistic T_L is applied to the *NYS LD* data, as well as LD data of Connecticut State, New Jersey State, and Pennsylvania State in this section. The reasons that these other three states have been chosen are: 1) They are geographical neighbors to NYS; 2) In these three states, LD was present and incidence rates (per 100,000 population) at county level have been recorded roughly over same period as the *NYS LD* data.

In the following section, descriptions are first given to the LD data for the three ‘neighboring’ states to NYS. The geographic relations of the three states to NYS are displayed in Figure 4. The time period from which the data for each state were available and the number of counties per state are summarized in Table 2. The results from the tests are also given.

LD data of Connecticut, New Jersey, and Pennsylvania

The time period during which yearly LD counts and incidence rates were available at the county level for all eight counties in Connecticut State was from 1991 to 2002, for all 21 counties in New Jersey State was from 1990 to 2000, and

for all 67 counties in Pennsylvania State was from 1990 to 2001 (Table 2).

Test results of LD data for the four states

The relationships between county sample mean, $\bar{Y}$, and its sample variance-mean ratio, $S^2/\bar{Y}$, for New York, Connecticut, New Jersey, and Pennsylvania are displayed in Figure 5a-d. The relationship between log-sample mean, $\log \bar{Y}$, and log-sample

variance, $\log S^2$, for the four states are displayed in Figure 6a-d. The results of the test T_L for each of the four states are summarized in Table 3.

Note that the T_L test is developed under the assumption that j observations of Y_i are identical independent distributions, which is not the case in the *NYS LD* data. Figure 7 shows the LD incidence curves of each county over the years from 1990 to 2000, with a small map of NYS to indicate geographic locations of the counties. For example,

1. Westchester County’s incidence rate, the green curve, was high in 1990, but decreased over time.

2. Putnam County's, the pink curve, was high in 1990, increased from 1990 to 1996, and has decreased since then.
3. Dutchess County's, the blue curve, was high in 1990 and kept increasing over time.
4. Columbia County's, the red curve, was very low in 1990 and then gradually increased from 1990 to 1995. It has increased very rapidly since 1996. In 2000, Columbia County had the highest LD incidence rate in the United States (CDC, 2002).
5. Rensselaer County's, the black curve, was low in 1990 and stayed the same till 1998. Then it increased slightly from 1998 to 2000.

The finding above indicated that LD occurrence in some of the NYS counties had strong time trends. To adjust for this violation to the independence assumption, the T_L test is also applied to the partial *NYSLD* data after taking out counties having significant time trends. Figure 8a displays the relationships between county sample mean, $\bar{Y}$, and its sample variance-mean ratio, $S^2/\bar{Y}$, and Figure 8b displays the relationship between log-sample mean, $\log \bar{Y}$, and log-sample variance, $\log S^2$, for the partial *NYSLD* data. The result of the test T_L for it is summarized in the row NY_p of Table 3.

Note that in Figure 5a-d, 6a-d, and 8a-b, the same axis scales are used for plots displaying relationship between $\bar{Y}$ and $S^2/\bar{Y}$, so are for plots displaying relationships between $\log \bar{Y}$ and $\log S^2$, for convenience of comparisons.

The scatter plots in Figure 5a-d and Figure 6a-d show that the relation between $\log S^2$ and $\log \bar{Y}$ has a much clearer linear form than the relation between $S^2/\bar{Y}$ and $\bar{Y}$ for all the four states. In contrast to Figure 5a and Figure 6a for the entire *NYSLD* data, Figure 8a and b show that NYS counties with significant time trends tend to have larger LD counts and sample variances than those without time trends.

The facts that p -values of T_L were close to zero and 95% Confidence Interval (95%CI) of

β_l did not include one for the data from all the four states and for the partial *NYSLD* data give strong evidence of over-dispersion to Poisson. The p -value of T_L for Connecticut was less than 0.05 but greater than 0.01. This may be caused by larger variation due to the fewer number of observations. The consistent test results of β_l for both the entire and the partial *NYSLD* data indicate that the T_L test is robust to time trend in Poisson data. Note that the values of β_l for New York, Connecticut, and New Jersey are very close to each other (ranging from 1.26 to 1.72).

Conclusion

In this paper, two simple, easy to implement tests (T_S and T_L) are proposed for assessment of simultaneous over-dispersion of a group of random variables to the Poisson model. These tests specify a relationship between the mean and variance. However, they do not require specification of the distribution under the alternative. The tests are easy to implement: the T_S is computed as the t -test from an WLS regression, and the T_L from an OLS regression. In this sense, the two tests can be given the name 'regression-based' tests.

Simulation experiments implemented in samples of small (5), moderate (11 and 50), and large (100) sizes shows that the empirical test size matches the nominal size well for T_L , but is unacceptably liberal for T_S for all experimented sample sizes, which suggests the log-transformation makes T_L less possible to break assumptions of linear regression. It is noted that the reference test T_F has unfit empirical test size when sample size is small but performs fine with moderate or bigger sample sizes. This may be due to the fact that the Fisher statistic is strictly correct for continuous variables, which becomes more realistic for Poisson as the sample size N increases. The power simulation experiments performed on T_L and T_F treat power as a function of dispersion parameter of alternative distributions. The empirical comparisons of power curves suggest that although both tests have adequate power even for small sample size, the power characteristic of T_L empirically superior to T_F in terms of sensitivity to degree of over-dispersion. This is especially true when over-dispersion is moderate.

a. sample size = 5

b. sample size = 11

c. sample size = 50

d. sample size = 100

- (Notes: 1. x-axis is v and y-axis is the percentage of reject H_0 for plots a-d.
2. $TL1 = T_L$ for $\alpha = 0.01$; $TF1 = T_F$ for $\alpha = 0.01$; $TL5 = T_L$ for $\alpha = 0.05$; $TF5 = T_F$ for $\alpha = 0.05$.)

Figure 2. The empirical power curve of T_L and T_F under alternative hypothesis H_1 : $Y_i \sim NB$

TEST FOR SPATIO-TEMPORAL COUNTS BEING POISSON

a. sample size = 5

b. sample size = 11

c. sample size = 50

d. sample size = 100

(notes: 1. x-axis is ω and y-axis is the percentage of reject H_0 for plots a-d.
2. $TL1 = T_L$ for $\alpha=0.01$; $TF1 = T_F$ for $\alpha=0.01$; $TL5 = T_L$ for $\alpha=0.05$; $TF5 = T_F$ for $\alpha=0.05$.)

Figure 3. The empirical power curve of T_L , and T_F under alternative hypothesis H_1 :
 $Y_i \sim \text{ZIP}$

Figure 4. Geographic location of the four states

The most commonly used probability models for discrete data are binomial (Bin), Poisson (Pois), and Negative Binomial (NB) (series 1) and their corresponding zero-inflated models: zero-inflated binomial (ZIBin), zero-inflated Poisson (ZIP), and zero-inflated Negative Binomial (ZINB) (series 2). Each of them has different flexibility to model over-dispersion to Poisson. Table 4 summarizes their variance-mean relationships (VMR) and relative over/under dispersion to Poisson.

Table 4 reveals that the Bin probability model only allows under-dispersion to Poisson, while the ZIP, NB, and ZINB probability models only allow over-dispersion to Poisson. Among these six probability models, ZIBin is the most flexible model. It allows all the three situations (under-dispersion, over-dispersion, and none) based on different relative values of ω and n .

On the other hand, after we assess the over or under-dispersion of a data set using the test T_L , different choices of probability models can be recommended based on different estimated values of β_i (Table 5). For example, if a test result indicates that the estimated β_i is statistically significantly greater than one ($\beta_i > 1$), probability models that allows over-

dispersion will be recommended such as ZIBin, ZIP, NB, or ZINB.

Although the motivation and essential theory of these tests exploits only the equality between mean and variance, this approach can be extended to tests of other relationships between mean and variance.

Equal observation points (N) for each variable are assumed in this study. Future research can be done by studying a group of variables with unequal observation points (i.e., I independent variables, each with N_i observation points). In this paper, when the linear regression is applied to mean and variance-mean ratio, a common regression coefficient (β_i) is assumed for the group of variables. In the future research, individual regression coefficient (β_{li}) can be given to each variable and Bayesian approaches can be used to estimate the parameters of interest.

Applications of the T_L test to the *NYSLD* and the *LD* data for Connecticut, New Jersey, and Pennsylvania suggest that the Poisson model is not statistically consistent with these count data and a natural alternative is the Negative binomial model. The fact that the values of β_i for New York, Connecticut, and New Jersey are close to each other (ranging from 1.26 to 1.72)

TEST FOR SPATIO-TEMPORAL COUNTS BEING POISSON

Table 2. Summary of time periods and locations studied

	NY	CT	NJ	PA
Starting year	1990	1991	1990	1990
Ending year	2000	2002	2000	2001
No. year	11	12	11	12
No.county	57	8	21	67

Table 3. Summary of the tests T_L on LD data for the four states

State	T_L	
	beta_1 (95%CI)	p-value
NY	1.70 (1.61, 1.79)	0.00
NY_p	1.34 (1.16, 1.51)	0.00
CT	1.72 (1.22, 3.23)	0.03
NJ	1.72 (1.40, 2.05)	0.00
PA	1.26 (1.10, 1.43)	0.00

Table 4. Variance-mean relationships of six commonly used probability models for discrete data and their relative over/under-dispersion to Poisson

series 1	Variance-Mean Relation	dispersion
Bin	$\text{var}(Y)/E(Y) = 1 - E(Y)/n$	under
Pois	$\text{var}(Y)/E(Y) = 1$	none
NB	$\text{var}(Y)/E(Y) = 1 + E(Y)/\nu$	over
series 2	Variance-Mean Relation	dispersion
ZIBin	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega - 1/n}{1 - \omega} E(\tilde{Y})$	dependent *
ZIP	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega}{1 - \omega} E(\tilde{Y})$	over
ZINB	$\frac{\text{var}(\tilde{Y})}{E(\tilde{Y})} = 1 + \frac{\omega + 1/\nu}{1 - \omega} E(\tilde{Y})$	over

* The over or under-dispersion to Poisson is dependent on the relative values between ω and n : if $\omega > 1/n$, then this model is over-dispersion to Poisson; if $\omega = 1/n$, then this model is neither over nor under-dispersion to Poisson; if $\omega < 1/n$, then this model is under-dispersion to Poisson.

a. New York

b. Connecticut

c. New Jersey

d. Pennsylvania

Figure 5. Scatter plots of $S^2 / \bar{Y}$ vs. $\bar{Y}$ for the four states
 (note: x-axis is $\bar{Y}$ and y-axis is $S^2 / \bar{Y}$ for plots a-d.)

TEST FOR SPATIO-TEMPORAL COUNTS BEING POISSON

a. New York

b. Connecticut

c. New Jersey

d. Pennsylvania

Figure 6. Scatter plots of $\log S^2$ vs. $\log \bar{Y}$ for the four states
(note: x-axis is $\log \bar{Y}$ and y-axis is $\log S^2$ for plots a-d.)

Figure 7. Fifty-seven NYS county annual incidence rates from 1990 to 2000

Figure 8a. Scatter plots of $S^2 / \bar{Y}$
vs. $\bar{Y}$ for the partial NYS Figure 8b. Scatter plots of $\log S^2$
vs. $\log \bar{Y}$ for the partial NYS

TEST FOR SPATIO-TEMPORAL COUNTS BEING POISSON

Table 5. Recommended probability models based on estimated coefficient for regression of log-sample variance on log-sample mean

est. Beta1	suggested models					
	Bin	ZIBin	Pois	ZIP	NB	ZINB
> 1		+		+	+	+
1		+	+			
< 1	+	+				

may suggest a general pattern of LD existing in the studied geographic area.

Results from the *NYSLD* data suggested that the new test statistic T_L seems robust to data with time trend in Poisson model. This is probably related to the fact that sums involved in the averages of individual Poissons are also Poissons. However, more systematic studies are needed before making any determinant conclusions.

References

- Bohning, D. (1994). A note on a test for Poisson overdispersion. *Biometrika* 81(2), 418-9.
- Brown, L. D., & Zhao, L. H. (2002). A test for the Poisson distribution. *Sankhya: The Indian Journal of Statistics* 64(A3), 611-25.
- Cameron, A. C., & Trivedi, P. K. (1990). Regression based tests for overdispersion in the Poisson model. *Journal of Econometrics* 46, 347-64.
- Centers for Disease Control and Prevention (CDC). (1997). Cases definitions for infectious conditions under public health surveillance. *Morbidity and Mortality Weekly Report* 46(RR-10), 1-55.
- Centers for Disease Control and Prevention (CDC) (2002). Summary of notifiable diseases—United States, 2000. *Morbidity and Mortality Weekly Report* 49(53), 1-16. Connecticut state department of health website (<http://www.dph.state.ct.us>), downloaded date: 06-11-04.
- Hedges, L. V., & Olkin, I. (1995). *Statistical Methods for Meta-Analysis*. Academic Press: Orlando, FL.
- Hinde, J., & Demetrio, C. (1998). Overdispersion: models and estimation. *Computational Statistics and Data Analysis* 27, 151-170.
- Hoel, P. G. (1943) On indices of dispersion. *Ann Math Statist* 14, 155-62.
- Johnson, N.L., Kotz, S., & Kemp A.W. (1992). *Univariate Discrete Distributions*, second edition. John Wiley & Sons Inc.: New York.
- Lambert, D., & Roeder, K. (1995). Overdispersion diagnostics for generalized linear models. *Journal of American Statistical Association* 90, 1225-36.
- Lindsey, J. K. (1995). *Modeling frequency and count data*. Clarendon Press: Oxford.
- New Jersey Department of Health and Senior Services website (<http://www.state.nj.us/health>), downloaded date: 06-11-04.
- Pennsylvania Department of Health website (<http://www.dsf.health.state.pa.us>), downloaded date: 06-11-04.
- Smyth, G. K. (2003). Pearson's goodness of fit statistic as a score test statistic. In: *Science and Statistics: A Festschrift for Terry Speed*. DR Goldstein (ed.). IMS Lecture Notes—Monograph Series, Volume 40, Institute of Mathematical Statistics, Beachwood, Ohio, pp115-26.
- Tiago de Oliveria, J. (1965). Some elementary tests of mixtures of discrete distributions. In: *Classical and Contagious Discrete Distributions*. GP Patil (ed.). Pergamon Press: New York, pp379-84.
- White, D. J., Chang, H-G., Benach, J. L., et al. (1991). Geographic spread and temporal increase of the Lyme disease epidemic. *JAMA* 266, 1230-6.

Two-Stage Short-Run (X, MR) Control Charts

Matthew E. Elam
Texas A&M University-Commerce

Kenneth E. Case
Oklahoma State University

This article is the first in a series of two articles that applies two-stage short-run control charting to (X, MR) charts. Theory is developed and then used to derive the control chart factor equations. In the sequel, the control chart factor calculations are computerized and an example is presented.

Key words: control chart, short-run statistical process control, two-stage control charting, probability integral of the range, probability integral of the studentized range, distribution of the mean moving range

Introduction

The statistical analysis of sample data often requires the sample to be random. In a random sample, each value comes from the same population distribution. Many situations exist in which it is difficult to obtain a random sample. One of these is when the population is not well-defined, as is the case when studying on-going processes, which are often encountered in manufacturing situations.

A statistical technique for establishing data as random in this situation is control charting. The upper and lower control limits and center line for control charts are constructed from data collected as some number m of subgroups, each having size n . Subgroup statistics are then plotted on the control charts. If these statistics plot between the control limits in a random pattern, then the data is likely random.

If not, a procedure is invoked to remove the offending subgroups to establish the data as random. The focus of this article is control charting in data limited (short-run) situations when using $n=1$.

Short-run control charting, as described by Hillier (1969), is necessary in the initiation of a new process, during the startup of a process just brought into statistical control again, and for a process whose total output is not large enough to use conventional control chart constants. Each of these is an example of a short-run situation. A short-run situation is one in which little or no historical information is available about a process in order to estimate process parameters to begin control charting. Consequently, the initial data obtained from the early run of the process must be used for this purpose.

When control charting in a short-run situation, Hillier (1969) gave a two-stage procedure that must be followed to set control limits that result in both the desired probability of a false alarm and a high probability of detecting a special cause signal. In the first stage, m initial subgroups of size n are drawn from the process and are used to determine the control limits. The initial subgroups are plotted against the control limits to retrospectively test if the process was in control while the initial subgroups were being drawn. Once control is established, the procedure moves to the second stage, where the subgroups that were not deleted in the first stage are used to determine the control limits for testing if the process remains in control while future subgroups are drawn. Each stage uses a different set of control chart

Matthew E. Elam is an Associate Professor of Industrial Engineering and Technology at Texas A&M University-Commerce and is an ASQ Certified Quality Engineer. Email him at Matthew_Elam@tamu-commerce.edu. Kenneth E. Case is Regents Professor Emeritus of Industrial Engineering and Management at Oklahoma State University. He is a fellow and past president of the ASQ and IIE, and is an Academician in the International Academy for Quality.

TWO-STAGE SHORT-RUN ($\bar{X}$, MR) CONTROL CHARTS

factors called first-stage short-run control chart factors and second-stage short-run control chart factors.

Hillier (1969) presented a two-stage short-run theory initially for $(\bar{X}, R)$ control charts (R is the range of a subgroup) and gave extensive results for first- and second-stage short-run control chart factors for $(\bar{X}, R)$ charts, but for $n=5$ only. Pyzdek (1993) and Yang (1995, 1999, 2000) attempted to expand Hillier's (1969) results for two-stage short-run $(\bar{X}, R)$ control charts, but their results contained incorrect values. Elam and Case (2001), as well as Elam (2001), described the development and execution of a computer program that overcomes the problems associated with Hillier's (1969), Pyzdek's (1993), and Yang's (1995, 1999, 2000) efforts to present two-stage short-run control chart factors for $(\bar{X}, R)$ charts.

The second application of Hillier's (1969) two-stage short-run theory was to $(\bar{X}, v)$ and $(\bar{X}, \sqrt{v})$ control charts (v is the variance of a subgroup). Yang and Hillier (1970) followed Hillier's (1969) theory to derive equations for calculating the factors required to determine two-stage short-run control limits for $(\bar{X}, v)$ and $(\bar{X}, \sqrt{v})$ charts. The tables of factors Yang and Hillier (1970) presented (see their Tables 1-6) were for several values for number of subgroups, α for the $\bar{X}$ chart, and α for the v and $\sqrt{v}$ charts both above the upper control limit and below the lower control limit (α is the probability of a false alarm). However, as in Hillier (1969), the results were for $n=5$ only. Elam and Case (2003a, 2003b) addressed issues concerning Yang and Hillier's (1970) results.

The third application of Hillier's (1969) two-stage short-run theory was to $(\bar{X}, s)$ control charts (s is the standard deviation of a subgroup). The difference between $(\bar{X}, \sqrt{v})$ and $(\bar{X}, s)$ control charts is that the former are constructed using the statistic $\sqrt{v}$ and the latter are constructed using the statistic $\bar{s}$. Elam and Case (2005a) developed the theory that was needed to apply Hillier's (1969) two-stage short-

run theory to $(\bar{X}, s)$ control charts. They then used this theory to derive the equations for calculating the factors required to determine two-stage short-run control limits for $(\bar{X}, s)$ charts. In a second article, Elam and Case (2005b) used the equations presented in Elam and Case (2005a) to develop a computer program that accurately calculates first- and second-stage short-run control chart factors for $(\bar{X}, s)$ charts regardless of the subgroup size, number of subgroups, α for the $\bar{X}$ chart, and α for the s chart both above the upper control limit and below the lower control limit.

Problem

It seems that no attempt appears in the literature to derive equations for calculating the factors required to determine two-stage short-run control limits for (X, MR) charts (MR is the moving range for two individual values). Del Castillo and Montgomery (1994) and Quesenberry (1995) both pointed out this deficiency. The application of (X, MR) control charts is desirable because, in a short-run situation, it may be difficult to form subgroups (Del Castillo & Montgomery, 1994).

Pyzdek (1993) attempted to present two-stage short-run control chart factors for (X, MR) charts for several values for number of subgroups and one value each for α for the X chart and α for the MR chart above the upper control limit. However, all of Pyzdek's (1993) Table 1 results for subgroup size one are incorrect because he used invalid theory (this is explained in detail in the Conclusion section).

Solution

First, the theory is developed that is needed to apply Hillier's (1969) two-stage short-run theory to (X, MR) control charts. It is then used to derive the equations for calculating the factors required to determine two-stage short-run control limits for (X, MR) charts. In the second article, Elam and Case (2006) used the equations presented in this article to develop a computer program that accurately calculates first- and second-stage short-run control chart factors for (X, MR) charts regardless of the number of subgroups, α for the X chart, and α for the MR

chart both above the upper control limit and below the lower control limit.

Outline

The probability integrals of the range and the studentized range are presented, both for subgroup size two. These are essential in the application of Hillier's (1969) theory to (X, MR) control charts. Next, Patnaik's (1950) theory is used to develop an approximation to the distribution of the mean moving range. From this result, equations for calculating the factors required to determine two-stage short-run control limits for (X, MR) charts are derived by following the work in the appendix of Hillier (1969). Also, equations to calculate conventional control chart constants for (X, MR) charts are derived. This article concludes with a discussion of its corrections to the literature.

Methodology

The Probability Integral of the Range for Subgroup Size Two

The probability integral (or cumulative distribution function (cdf)) of the range for subgroups of size two sampled from a standard Normal population was given by Pachares (1959) as equation (1) (with some modifications in notation):

$$P(W) = 2 \times \int_{-\infty}^{\infty} f(x) \times (F(x + W) - F(x)) dx \quad (1)$$

W represents the (standardized) range w/σ , where w is the range of a subgroup and σ is the population standard deviation. Throughout this article, $F(x)$ is the cdf of the standard Normal probability density function (pdf) $f(x)$.

The mean of the distribution of the range $W = (w/\sigma)$ for subgroups of size two sampled from a Normal population with mean μ and variance equal to one given by Harter (1960) is equation (2) (with some modifications in notation):

$$d2 = 2 / \pi^{0.5} \quad (2)$$

The value $d2$ is the control chart constant denoted by d_2 (see Table M in the appendix of

Duncan, 1974). The equation for $d2$ for subgroup size two for any value of σ was given by Johnson, Kotz, and Balakrishnan (1994). Equations (1) and (2) are the forms used in the computer program in Elam and Case (2006).

The Probability Integral of the Studentized Range for Subgroup Size Two

The probability integral of the studentized range for subgroups of size two sampled from a Normal population was given by Harter, Clemm, and Guthrie (1959) as equation (3a):

$$P3(z) = (5/z) \times \exp(cv) \times (P1(z) + P2(z)) \quad (3a)$$

where

$$cv = \ln(2) + (v/2) \times \ln(v/2) - (v/2) - \text{gammln}(v/2) \quad (3b)$$

$$P1(z) = \int_0^{11} [5 \times (W/z) \times \exp((z^2 - 25 \times W^2)/(2 \times z^2))]^{v-1} \times \exp((z^2 - 25 \times W^2)/(2 \times z^2)) \times P(W) dW \quad (3c)$$

$$P2(z) = (z/5) \times \int_{55/z}^{\infty} (x \times \exp((1 - x^2)/2))^{v-1} \times \exp((1 - x^2)/2) dx \quad (3d)$$

The variable z is equal to $5 \times Q$. Q represents the studentized range w/s , where w is the range of a subgroup and s is an independent estimate (based on v degrees of freedom) of the population standard deviation. The equation for determining v is derived in the next subsection. The equation for cv (equation (3b)) is the natural logarithm of the equation for $C(v)$ given by Harter, Clemm, and Guthrie (1959). It is derived in Appendix I: Derivations of Elam and Case (2001). The function gammln represents the natural logarithm of the gamma (Γ) function. In equation (3c), $P(W)$ is the probability integral of the range $W = (w/\sigma)$ for subgroup size two (see equation (1)). Equations (3a)-(3d) are the forms used in the computer program in Elam and Case (2006) because they allow for large values of v

TWO-STAGE SHORT-RUN (X, MR) CONTROL CHARTS

(hence large values for m (the number of subgroups)) in the program.

As $v \rightarrow \infty$ (i.e., as $m \rightarrow \infty$), the distribution of the studentized range $Q = (w/s)$ for subgroup size two converges to the distribution of the range $W = (w/\sigma)$ for subgroup size two (see Pearson and Hartley, 1943). This fact is used to derive equations to calculate α -based conventional control chart constants for the MR chart.

The Distribution of the Mean Moving Range

Consider the situation in which the mean of a statistic is calculated by averaging m values of the statistic, each of which is calculated from a subgroup of size n . Patnaik (1950) investigated this situation when the statistic was the range and developed an approximation to the distribution of the mean range $\bar{R}/\sigma$. The resulting distribution was the $(\chi \times d_2^*)/\sqrt{v}$ distribution, which is a function of the χ distribution with v degrees of freedom (the χ distribution with v degrees of freedom and its moments about zero may be found in Johnson and Welch, 1939).

Equations for v and d_2^* were derived from results obtained by equating the squared means as well as the variances of the distribution of the mean range $\bar{R}/\sigma$ and the $(\chi \times d_2^*)/\sqrt{v}$ distribution with v degrees of freedom. Hillier (1964, 1967) used Patnaik's (1950) theory to derive equations to calculate short-run control chart factors for $\bar{X}$ and R charts, respectively. Hillier (1969) then incorporated the two-stage procedure into his short-run control chart factor calculations for $(\bar{X}, R)$ charts.

Consider the situation in which the statistic is the moving range of size two and the distribution of interest is the distribution of the mean moving range $\overline{MR}/\sigma$. Evidence exists in the literature that $\overline{MR}/\sigma$ may be approximated by a distribution that is a function of either the χ^2 or the χ distribution. Sathe and Kamat (1957) used results given by Cadwell (1953, 1954) to approximate the distribution of the mean successive difference (i.e., the distribution

of the mean moving range $\overline{MR}/\sigma$) by a distribution that is a function of a power of the χ^2 distribution. Roes, Does, and Schurink (1993) used theory similar to Patnaik's (1950) theory to approximate the distribution of the mean moving range $\overline{MR}/\sigma$ (with $\sigma=1.0$) by a distribution that is a function of the χ distribution.

In order to be able to use Hillier's (1969) theory to derive equations for calculating the factors required to determine two-stage short-run control limits for (X, MR) charts, Patnaik's (1950) theory was applied to approximate the distribution of the mean moving range $\overline{MR}/\sigma$ by the $(\chi \times d_2^*(MR))/\sqrt{v}$ distribution with v degrees of freedom (this v is the same as the one that appears in equation (3a)). The equation for $d_2^*(MR)$ is derived in the Appendix and is given as equation (4) (note: $d2starMR \equiv d_2^*(MR)$):

$$d2starMR = (d2^2 + d2^2 \times r)^{0.5} \quad (4)$$

The equation for the control chart constant $d2$ for subgroup size two is given earlier as equation (2). The value r represents the variance of $\overline{MR}/d2$. Its equation is given later as equation (7a). Equation (4) is the form used in the computer program in Elam and Case (2006).

Using results from Prescott (1971), the equation for v is determined by equating the ratio of the variance to the squared mean, both of the χ distribution with v degrees of freedom, to the ratio of the variance to the squared mean, both of the distribution of the mean moving range $\overline{MR}/\sigma$. The resulting equation for v is equation (5):

$$d(x) = h(x) - r \quad (5)$$

The exact value for v is the value of x such that $d(x)$ is equal to zero. The function $h(x)$ is the ratio of the variance to the squared mean, both of the χ distribution with x degrees of freedom (x replaces v). The mean and variance of the χ distribution with v degrees of freedom are given

in the Appendix. The equation for $h(x)$, which is derived in Appendix I: Derivations of Elam and Case (2001), is given as equation (6):

$$h(x) = (x \times \exp(2 \times (\text{gammln}(0.5 \times x) - \text{gammln}(0.5 \times x + 0.5))) - 2) / 2 \quad (6)$$

The value r is the ratio of the variance to the squared mean, both of the distribution of the mean moving range $\overline{MR}/\sigma$. The mean and the variance of the distribution of the mean moving range $\overline{MR}/\sigma$ are derived in the Appendix. The equation for r was given by Palm and Wheeler (1990) as equation (7a):

$$r = (b \times (m - 1) - c) / (m - 1)^2 \quad (7a)$$

where

$$b = 2 \times \pi / 3 - 3 + 3^{0.5} \quad (7b)$$

$$c = \pi / 6 - 2 + 3^{0.5} \quad (7c)$$

Cryer and Ryan (1990) gave an equivalent form for equation (7a). Hoel (1946) gave an equation for the variance of $\overline{MR}$ which, when multiplied by $1/d_2^2$, gives the same results as those obtained by using equation (7a). It should be noted that an equivalent form (also based on Patnaik's (1950) theory) of equation (5) may be found in Palm and Wheeler (1990), who used their result to calculate equivalent degrees of freedom for population standard deviation estimates based on consecutive overlapping moving ranges of size two. Equations (5), (6), and (7a)-(7c) are the forms used in the computer program in Elam and Case (2006).

Approximating the distribution of the mean moving range $\overline{MR}/\sigma$ by the $(\chi \times d_2^*(MR)) / \sqrt{v}$ distribution with v degrees of freedom works well. In fact, based on how $d_2^*(MR)$ is derived in the Appendix, the means and variances of these two distributions are equal.

Results

Because the $(\chi \times d_2^*(MR)) / \sqrt{v}$ distribution with v degrees of freedom approximates the distribution of the mean moving range $\overline{MR}/\sigma$, the derivation of equations to calculate first- and second-stage short-run control chart factors for (X, MR) charts follows the work in the appendix of Hillier (1969). E22, the second-stage short-run control chart factor for the X chart, is derived in almost the same manner as Hillier's (1969) A_2^* . Differences are that $n=1$ and X , $\overline{X}$, $\overline{MR}$, and $d_2^*(MR)$ in this article replace $\overline{X}$, $\overline{X}$, A_2^* , $\overline{R}$, and c , respectively, in Hillier (1969). The resulting equation for E22 is given as equation (8) (note: $d2starMR \equiv d_2^*(MR)$):

$$E22 = (\text{crit_t} / d2starMR) \times ((m + 1) / m)^{0.5} \quad (8)$$

The value crit_t is the critical value for a cumulative area of $1 - (\alpha_{\text{Ind}}/2)$ under the Student's t curve with v degrees of freedom (α_{Ind} is the probability of a false alarm on the X control chart). Equation (8) is the form used in the computer program in Elam and Case (2006).

E21, the first-stage short-run control chart factor for the X chart, is derived in almost the same manner as Hillier's (1969) A_2^{**} . Differences are that E21, X_i , $\overline{X}$, $\overline{MR}$, and $d_2^*(MR)$ in this article replace A_2^{**} , $\overline{X}_i$, $\overline{X}$, $\overline{R}$, and c , respectively, in Hillier (1969). The resulting equation for E21 is given as equation (9):

$$E21 = (\text{crit_t} / d2starMR) \times ((m - 1) / m)^{0.5} \quad (9)$$

The value crit_t has the same meaning here as in equation (8). Equation (9) is the form used in the computer program in Elam and Case (2006).

D42, the second-stage short-run upper control chart factor for the MR chart, is derived in the Appendix. Other than differences in notation, this derivation follows that for Hillier's

TWO-STAGE SHORT-RUN (X, MR) CONTROL CHARTS

(1969) D_4^* . The resulting equation for D42 is given as equation (10):

$$D42 = qD4 / d2starMR \quad (10)$$

The value $qD4$ is the $1-\alpha$ MRUCL percentage point of the distribution of the studentized range $Q = (w/s)$ for subgroup size two with v degrees of freedom (α MRUCL is the probability of a false alarm on the MR chart above the upper control limit). Equation (10) is the form used in the computer program in Elam and Case (2006).

D32, the second-stage short-run lower control chart factor for the MR chart, is derived in a manner similar to D42. Differences are that D32, $qD3$, and α MRLCL replace D42, $qD4$, and $1-\alpha$ MRUCL, respectively (α MRLCL is the probability of a false alarm on the MR chart below the lower control limit). The resulting equation for D32 is given as equation (11):

$$D32 = qD3 / d2starMR \quad (11)$$

The value $qD3$ is the α MRLCL percentage point of the distribution of the studentized range $Q = (w/s)$ for subgroup size two with v degrees of freedom. Equation (11) is the form used in the computer program in Elam and Case (2006).

D41, the first-stage short-run upper control chart factor for the MR chart, is derived in almost the same manner as Hillier's (1969) D_4^{**} . Differences are that D41, MR_i , D42, and $\overline{MR}$ in this paper replace D_4^{**} , R_i , D_4^* , and $\overline{R}$, respectively, in Hillier (1969). D41 is given as equation (12):

$$D41 = m \times qD4prevm / (d2starMRprevm \times (m-1) + qD4prevm) \quad (12)$$

The value $qD4prevm$ is the $1-\alpha$ MRUCL percentage point of the distribution of the studentized range $Q = (w/s)$ for subgroup size two with $vprevm$ degrees of freedom (the value

$vprevm$ has the same meaning as v , except it is for $m-1$ subgroups). The value $d2starMRprevm$ has the same equation as $d2starMR$ (given earlier as equation (4)), except m is replaced with $m-1$. Equation (12) is the form used in the computer program in Elam and Case (2006).

The equation for D31, the first-stage short-run lower control chart factor for the MR chart, is derived in almost the same manner as Hillier's (1969) D_3^{**} . Differences are that D31, MR_i , D32, and $\overline{MR}$ in this article replace D_3^{**} , R_i , D_3^* , and $\overline{R}$, respectively, in Hillier (1969). The resulting equation for D31 is given as equation (13):

$$D31 = m \times qD3prevm / (d2starMRprevm \times (m-1) + qD3prevm) \quad (13)$$

The value $qD3prevm$ is the α MRLCL percentage point of the distribution of the studentized range $Q = (w/s)$ for subgroup size two with $vprevm$ degrees of freedom. Equation (13) is the form used in the computer program in Elam and Case (2006).

The equation for E2, the conventional control chart constant for the X chart, may be obtained by taking the limit of either E21 or E22 as $m \rightarrow \infty$ (i.e., as $v \rightarrow \infty$). The resulting equation for E2 is given as equation (14):

$$E2 = \text{crit_}z / d2 \quad (14)$$

The value $\text{crit_}z$ is the critical value for a cumulative area of $1-(\alpha \ln d/2)$ under the standard Normal curve. The equation for the control chart constant $d2$ for subgroup size two is given earlier as equation (2). Equation (14) is the form used in the computer program in Elam and Case (2006).

The equation for D4, the α -based conventional upper control chart constant for the MR chart, may be obtained by taking the limit of either D41 as $m \rightarrow \infty$ (i.e., as $vprevm \rightarrow \infty$) or D42 as $m \rightarrow \infty$ (i.e., as $v \rightarrow \infty$). The resulting equation for D4 is given as equation (15):

$$D4 = wD4 / d2 \quad (15)$$

The value $wD4$ is the $1-\alpha$ MRUCL percentage point of the distribution of the range $W = (w/\sigma)$ for subgroup size two. Equation (15) is the form used in the computer program in Elam and Case (2006).

The equation for $D3$, the α -based conventional lower control chart constant for the MR chart, may be obtained by taking the limit of either $D31$ as $m \rightarrow \infty$ (i.e., as $v_{prevm} \rightarrow \infty$) or $D32$ as $m \rightarrow \infty$ (i.e., as $v \rightarrow \infty$). The resulting equation for $D3$ is given as equation (16):

$$D3 = wD3/d2 \quad (16)$$

The value $wD3$ is the α MRLCL percentage point of the distribution of the range $W = (w/\sigma)$ for subgroup size two. Equation (16) is the form used in the computer program in Elam and Case (2006).

Conclusion

As mentioned in the Problem subsection of the Introduction, all of Pyzdek's (1993) Table 1 results for subgroup size one are incorrect because he used invalid theory. This is true for two reasons. The first is that he used degrees of freedom based on Patnaik's (1950) approximation applied to the distribution of the mean range $\bar{R}/\sigma$, where $\bar{R}$ is the average of m values of R , each based on a subgroup of size two, not the distribution of the mean moving range $\overline{MR}/\sigma$. In the latter case, the degrees of freedom reflect the fact that serial correlation exists among consecutive overlapping moving ranges of size two, which means that the average of these overlapping MRs reflects that serial correlation. The result is that degrees of freedom based on Patnaik's (1950) approximation applied to the distribution of the mean moving range $\overline{MR}/\sigma$ is less than that from applying Patnaik's (1950) approximation to the distribution of the mean range $\bar{R}/\sigma$, where R is the range of a subgroup of size two.

The second is that Pyzdek (1993) used the equation for d_2^* (i.e., $d2star$) instead of that

for $d2starMR$ (given earlier as equation (4)). The equation for d_2^* is given as equation (17):

$$d_2^* = (d_2^2 + d_3^2 / m)^{0.5} \quad (17)$$

where d_2 and d_3 are the mean and standard deviation, respectively, of the distribution of the range $W = (w/\sigma)$. Equations to calculate d_2 and d_3 for any subgroup size as well as the equation for d_2^* may be found in Elam and Case (2001). The difference between equations (4) and (17) is that equation (4) has $d2^2 \times r$, which is the variance of the distribution of the mean moving range $\overline{MR}/\sigma$, instead of d_3^2/m , which is the variance of the distribution of the mean range $\bar{R}/\sigma$. The equation for r in $d2^2 \times r$ reflects the fact that serial correlation exists among consecutive overlapping moving ranges of size two, which means that the average of these overlapping MRs reflects that serial correlation. The result is that values for $d2starMR$ are less than those for $d2star$ for subgroup size two; but, as $m \rightarrow \infty$, both converge to $d2$. It should be noted that $d2starMR$ for $m=2$ is equal to $d2star$ for $n=2$ and $m=1$ (see Table A1 in Appendix III: Tables of Elam and Case, 2001).

One last issue regarding Pyzdek's (1993) Table 1 results is that he gave second-stage short-run control chart factors for number of subgroups equal to one. This is impossible because one must have two subgroups in order to calculate one moving range. For first-stage short-run control chart factors for the individuals and moving range charts, m must be at least two and three, respectively. The reason is $E21$ (see equation (9)) includes $d2starMR$ (see equation (4)), which includes r , which, according to equation (7a), must have m at least two. Also, in equations (12) and (13), $D41$ and $D31$, respectively, include $d2starMR_{prevm}$, which includes r_{prevm} (r for $m-1$ subgroups), which must have m at least three. For second-stage short-run control chart factors for the individuals and moving range charts, m must be at least two.

TWO-STAGE SHORT-RUN (X, MR) CONTROL CHARTS

References

- Cadwell, J. H. (1953). Approximating to the distributions of measures of dispersion by a power of χ^2 . *Biometrika*, 40, 336-346.
- Cadwell, J. H. (1954). The statistical treatment of mean deviation. *Biometrika*, 41, 12-18.
- Cryer, J. D. & Ryan, T. P. (1990). The estimation of sigma for an X chart: $\overline{MR}/d_2$ or S/c_4 ?. *Journal of Quality Technology*, 22(3), 187-192.
- Del Castillo, E. & Montgomery, D. C. (1994). Short-run statistical process control: Q-chart enhancements and alternative methods. *Quality and Reliability Engineering International*, 10, 87-97.
- Duncan, A. J. (1974). *Quality control and industrial statistics*. 4th ed. Homewood, IL: Richard D. Irwin, Inc.
- Elam, M. E. (2001). *Investigation, extension, and generalization of a methodology for two stage short run variables control charting*. Ph.D. dissertation. Stillwater, Oklahoma: Oklahoma State University.
- Elam, M. E. & Case, K. E. (2001). A computer program to calculate two-stage short-run control chart factors for $(\bar{X}, R)$ charts. *Quality Engineering*, 14(1), 77-102.
- Elam, M. E. & Case, K. E. (2003a). Two-stage short-run $(\bar{X}, v)$ and $(\bar{X}, \sqrt{v})$ control charts. *Quality Engineering*, 15(3), 441-448.
- Elam, M. E. & Case, K. E. (2003b). A computer program to calculate two-stage short-run control chart factors for $(\bar{X}, v)$ and $(\bar{X}, \sqrt{v})$ charts. *Quality Engineering*, 15(4), 609-638.
- Elam, M. E. & Case, K. E. (2005a). Two-stage short-run $(\bar{X}, s)$ control charts. *Quality Engineering*, 17(1), 95-107.
- Elam, M. E. & Case, K. E. (2005b). A computer program to calculate two-stage short-run control chart factors for $(\bar{X}, s)$ charts. *Quality Engineering*, 17(2), 259-277.
- Elam, M. E. & Case, K. E. (2006). A computer program to calculate two-stage short-run control chart factors for (X, MR) charts. *Journal of Statistical Software*, 15(11), <http://www.jstatsoft.org/>.
- Harter, H. L. (1960). Tables of range and studentized range. *Annals of Mathematical Statistics*, 31, 1122-1147.
- Harter, H. L., Clemm, D. S., & Guthrie, E. H. (1959). *The probability integrals of the range and the studentized range – probability integral and percentage points of the studentized range; critical values for Duncan's new multiple range test*. Wright Air Development Center Technical Report 58-484, Vol. II. Defense Technical Information Center (DTIC) Document AD Number 231733.
- Hillier, F. S. (1964). $\bar{X}$ chart control limits based on a small number of subgroups. *Industrial Quality Control*, 20(8), 24-29.
- Hillier, F. S. (1967). Small sample probability limits for the range chart. *Journal of the American Statistical Association*, 62(320), 1488-1493.
- Hillier, F. S. (1969). $\bar{X}$ - and R-chart control limits based on a small number of subgroups. *Journal of Quality Technology*, 1(1), 17-26.
- Hoel, P. G. (1946). The efficiency of the mean moving range. *Annals of Mathematical Statistics*, 17, 475-482.
- Johnson, N. L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions - volume 1*. 2nd ed. New York, NY: John Wiley and Sons, Inc.
- Johnson, N. L. & Welch, B. L. (1939). On the calculation of the cumulants of the χ -distribution. *Biometrika*, 31, 216-218.
- Pachares, J. (1959). Table of the upper 10% points of the studentized range. *Biometrika*, 46, 461-466.
- Palm, A. C. & Wheeler, D. J. (1990). Equivalent degrees of freedom for estimates of the process standard deviation based on Shewhart control charts. Unpublished Manuscript.

Patnaik, P. B. (1950). The use of mean range as an estimator of variance in statistical tests. *Biometrika*, 37, 78-87.

Pearson, E. S. & Hartley, H. O. (1943). Tables of the probability integral of the studentized range. *Biometrika*, 33, 89-99.

Prescott, P. (1971). Use of a simple range-type estimator of σ in tests of hypotheses. *Biometrika*, 58(2), 333-340.

Pyzdek, T. (1993). Process control for short and small runs. *Quality Progress*, 26(4), 51-60.

Quesenberry, C. P. (1995). Response. *Journal of Quality Technology*, 27(4), 333-343.

Roes, K. C. B., Does, R. J. M. M., & Schurink, Y. (1993). Shewhart-type control charts for individual observations. *Journal of Quality Technology*, 25(3), 188-198.

Sathe, Y. S. & Kamat, A. R. (1957). Approximations to the distributions of some measures of dispersion based on successive differences. *Biometrika*, 44, 349-359.

Yang, S.-F. (1995). *Control policy for small and short production runs*. Taiwan, ROC: National Science Council Research Report.

Yang, S.-F. (1999). An approach to controlling process variability for short production runs. *Total Quality Management*, 10(8), 1123-1129.

Yang, S.-F. (2000). Statistical process control for short run manufacturing systems. *Process Control and Quality*, 11(5), 433-439.

Yang, C.-H. & Hillier, F. S. (1970). Mean and variance control chart limits based on a small number of subgroups. *Journal of Quality Technology*, 2(1), 9-16.

Appendix

Derive:

$$d2starMR = (d2^2 + d2^2 \times r)^{0.5}$$

First, the mean and variance of the distribution of the mean moving range $\overline{MR}/\sigma$ need to be determined. Note: By definition,

$$E(MR/\sigma) = d2 \Rightarrow (1/\sigma) \times E(MR) = d2 \Rightarrow E(MR) = d2 \times \sigma$$

$$\begin{aligned} E(\overline{MR}/\sigma) &= (1/\sigma) \times E(\overline{MR}) \\ &= (1/\sigma) \times E\left(\sum_{i=1}^{m-1} MR_i / (m-1)\right) \\ &= (1/\sigma) \times (1/(m-1)) \times E\left(\sum_{i=1}^{m-1} MR_i\right) \\ &\Rightarrow E(\overline{MR}/\sigma) \\ &= (1/\sigma) \times (1/(m-1)) \times \sum_{i=1}^{m-1} E(MR_i) \\ &= (1/\sigma) \times (1/(m-1)) \times \sum_{i=1}^{m-1} (d2 \times \sigma) \end{aligned}$$

because

$$E(MR) = d2 \times \sigma.$$

$$\Rightarrow E(\overline{MR}/\sigma) = (1/\sigma) \times (1/(m-1)) \times ((m-1) \times d2 \times \sigma) = d2$$

$$\text{Var}(\overline{MR}/\sigma) = (1/\sigma^2) \times \text{Var}(\overline{MR})$$

From Palm and Wheeler (1990),

$$\text{Var}(\overline{MR}/d2) = \sigma^2 \times r$$

where

$$r = (b \times (m-1) - c) / (m-1)^2,$$

with

$$b = 2 \times \pi / 3 - 3 + \sqrt{3}$$

and

$$c = \pi / 6 - 2 + \sqrt{3}$$

$$\begin{aligned} \Rightarrow r &= (1/\sigma^2) \times \text{Var}(\overline{MR}/d2) \\ &= (1/\sigma^2) \times (1/d2^2) \times \text{Var}(\overline{MR}) \end{aligned}$$

TWO-STAGE SHORT-RUN (X, MR) CONTROL CHARTS

$$\begin{aligned} &\Rightarrow d2^2 \times r \\ &= (1/\sigma^2) \times \text{Var}(\overline{MR}) \\ &\Rightarrow \text{Var}(\overline{MR}/\sigma) \\ &= d2^2 \times r \end{aligned}$$

According to Johnson and Welch (1939), the mean of the χ distribution with ν degrees of freedom is calculated using the following equation (with some modifications in notation):

$$\begin{aligned} E(\chi) &= \sqrt{2} \times \Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu) \\ &\Rightarrow E(\chi \times d2starMR / \sqrt{\nu}) \\ &= (d2starMR / \sqrt{\nu}) \times E(\chi) \\ &= \sqrt{2} \times (d2starMR / \sqrt{\nu}) \\ &\quad \times (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu)) \end{aligned}$$

Equating the squared means of the distribution of the mean moving range $\overline{MR}/\sigma$ and the $(\chi \times d2starMR)/\sqrt{\nu}$ distribution with ν degrees of freedom results in the following:

$$\begin{aligned} d2^2 &= 2 \times (d2starMR^2 / \nu) \\ &\quad \times (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu))^2 \\ &\Rightarrow d2starMR^2 = d2^2 \times (\nu / 2) \\ &\quad \times (\Gamma(0.5 \times \nu) / \Gamma(0.5 \times \nu + 0.5))^2 \quad (A.1) \end{aligned}$$

Appendix 7 of Elam and Case (2005a) gave the variance of the χ distribution with ν degrees of freedom as follows:

$$\text{Var}(\chi) = \nu - 2 \times (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu))^2$$

$$\begin{aligned} &\Rightarrow \text{Var}(\chi \times d2starMR / \sqrt{\nu}) \\ &= (d2starMR^2 / \nu) \times \text{Var}(\chi) \\ &= (d2starMR^2 / \nu) \\ &\quad \times (\nu - 2 \times (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu))^2) \end{aligned}$$

Equating the variances of the distribution of the mean moving range $\overline{MR}/\sigma$ and the $(\chi \times d2starMR)/\sqrt{\nu}$ distribution with ν degrees of freedom results in the following:

$$\begin{aligned} d2^2 \times r &= (d2starMR^2 / \nu) \\ &\quad \times (\nu - 2 \times (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu))^2) \\ &\Rightarrow (\Gamma(0.5 \times \nu + 0.5) / \Gamma(0.5 \times \nu))^2 \\ &= (d2^2 \times r \times \nu / d2starMR^2 - \nu) / (-2) \\ &\Rightarrow (\Gamma(0.5 \times \nu) / \Gamma(0.5 \times \nu + 0.5))^2 \\ &= 2 / (\nu \times (1 - d2^2 \times r / d2starMR^2)) \quad (A.2) \end{aligned}$$

Substituting equation (A.2) into equation (A.1) gives the following equation:

$$\begin{aligned} d2starMR^2 &= d2^2 \times (\nu / 2) \\ &\quad \times (2 / (\nu \times (1 - d2^2 \times r / d2starMR^2))) \\ &\Rightarrow d2starMR^2 \\ &= d2^2 / (1 - d2^2 \times r / d2starMR^2) \\ &= d2starMR^2 \times d2^2 / (d2starMR^2 - d2^2 \times r) \end{aligned}$$

$$\Rightarrow 1 = d2^2 / (d2starMR^2 - d2^2 \times r)$$

$$\Rightarrow d2starMR^2 = d2^2 + d2^2 \times r$$

$$\Rightarrow d2starMR = (d2^2 + d2^2 \times r)^{0.5}$$

Derive:

$D_{42} = (qD_4 / d_{2\text{star}} \overline{MR})$, where qD_4 is the $1 - \alpha$ MRUCL percentage point of the distribution of the studentized range $Q = (w/s)$ for subgroup size two with v degrees of freedom (α MRUCL is the probability of a false alarm on the MR chart above the upper control limit).

Notes: The ensuing derivation is based on the derivation of D_4^* in the appendix of Hillier (1969). The value MR denotes the moving range of a subgroup of size two drawn while in the second stage of the two-stage procedure.

The value D_{42} needs to be determined such that the following holds:

$$\begin{aligned} P(MR \leq D_{42} \times \overline{MR}) &= 1 - \alpha \text{MRUCL} \\ \Rightarrow P(MR / \overline{MR} \leq D_{42}) &= 1 - \alpha \text{MRUCL} \end{aligned}$$

To do this, the probability distribution of $MR / \overline{MR}$ needs to be determined. Notice that MR / σ is the statistic for the distribution of the range $W = (w/\sigma)$ for subgroup size two. An independent estimate of σ based on $\overline{MR}$ is now needed. Replacing σ with this independent estimate results in the statistic for the distribution of the studentized range $Q = (w/s)$ for subgroup size two, which has v degrees of

freedom. The equation to calculate v is based on the fact that the Patnaik (1950) approximation has been applied to the distribution of the mean moving range. As a result, $\overline{MR} / d_{2\text{star}} \overline{MR}$ needs to be used.

$$\begin{aligned} &\Rightarrow MR / \sigma \\ &= MR / (\overline{MR} / d_{2\text{star}} \overline{MR}) \\ &= MR \times d_{2\text{star}} \overline{MR} / \overline{MR} \end{aligned}$$

where

$$(MR \times d_{2\text{star}} \overline{MR}) / \overline{MR}$$

is the statistic for the distribution of the studentized range $Q = (w/s)$ for subgroup size two with v degrees of freedom.

$$\begin{aligned} &\Rightarrow 1 - \alpha \text{MRUCL} \\ &= P(MR \times d_{2\text{star}} \overline{MR} / \overline{MR} \leq qD_4) \\ &= P(MR / \overline{MR} \leq qD_4 / d_{2\text{star}} \overline{MR}) \end{aligned}$$

Setting

$$\begin{aligned} D_{42} &= qD_4 / d_{2\text{star}} \overline{MR} \Rightarrow 1 - \alpha \text{MRUCL} \\ &= P(MR / \overline{MR} \leq D_{42}) = P(MR \leq D_{42} \times \overline{MR}) \end{aligned}$$

Measuring Overall Heterogeneity in Meta-Analyses: Application to CSF Biomarker Studies in Alzheimer's Disease

Chengjie Xiong¹, Feng Gao¹, Yan Yan^{1,2}, Jingqin Luo¹, YunJu Sung¹, & Gang Shi¹
Washington University in St. Louis

The interpretations of statistical inferences from meta-analyses depend on the degree of heterogeneity in the meta-analyses. Several new indices of heterogeneity in meta-analyses are proposed, and assessed the variation/difference of these indices through a large simulation study. The proposed methods are applied to biomarkers of Alzheimer's disease.

Keywords: Alzheimer's disease, heterogeneity, meta-analysis, standard errors, uncertainty interval.

Introduction

Medical practitioners and their patients make decisions within the context of a rapidly changing body of scientific evidence on medicine and health care system that influence the availability, accessibility, and cost of diagnostic tests and therapies (Sackett & Haynes, 1995).

Timely, useful evidence from the biomedical literature should be an integral component of clinical and medical decision making. The importance of basing medical practice more firmly on the results of existing scientific evidence through systematic reviews was starkly demonstrated by a paper in the early 1990s (Antman, Kupelnick, Mosteller, & Chalmers, 1992), which compared the results of meta-analyses of trials of treatments for people who have suffered a heart attack as the trials were published with the recommendations of experts published in review articles and textbooks over the same time period.

This showed a significant divergence between expert recommendations and the summaries of the trials.

Ineffective treatments were being recommended, and highly effective treatments were not. As a result, lives that could have been saved were lost, and resources were wasted. Systematic reviews can be very useful medical decision-making tools by objectively summarizing large amounts of information, identifying gaps in medical research and evidence, and identifying beneficial or harmful interventions. Clinicians can use systematic reviews to guide their patient care. Consumers and patients can use systematic reviews to help them make health care decisions. Policymakers can use systematic reviews to help them make decisions about the types of health care to provide.

Systematic reviews can provide convincing and reliable evidence relevant to many aspects of medical and biological research and health care (Egger & Smith, 1997), especially when the results of individual studies they include show clinically important effects of comparable magnitude. Such reviews aim to comprehensively identify and assess all studies relevant to a given scientific question, and meta-analysis has been the major statistical methodology for the quantitative synthesis of study results. Many methods for meta-analysis are available, and the most popularly applied in the medical research focus on the optimum combination of published summary statistics in some form of weighted averages (DerSimonian

¹Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, U.S.A. ² Department of Surgery, Washington University in St. Louis, St. Louis, MO 63110, U.S.A. E-mail format: name@wubios.wustl.edu: chengjie@; feng@; yan@; rosy@; yunju@; gang@. All authors have research interests in Biostatistics and Bioinformatics.

& Laird, 1986; Egger, Smith, & Phillips, 1997; & Whitehead & Whitehead, 1991).

Usually, each study is given a weight according to the precision of its results on summary statistics. Studies with good precisions are weighted more heavily than studies with greater uncertainty. The variance for the overall estimate of the parameter under study in meta-analyses is in general from two different sources, one is associated with the individual studies (i.e., the within-study variance), and the other is associated with the possible difference between different studies (i.e., between-study variance). When the between-study variance is assumed to be 0, each study is simply weighted according to its own variance. This approach characterizes a fixed effects model which is exemplified by the Mantel-Haenszel method (Mantel & Haenszel, 1959) or the Peto method (Yusuf, Peto, Lewis, Collins, & Sleight, 1985).

When the between-study variance is not zero, methods which incorporate a between-study component of variation for the overall effect under estimation are based on random effects models (Laird, & Mosteller, 1990). The between-study variance represents the excessive variation in observed individual study effects over that expected from the imprecision of results within each study. Fixed effects and random effects model for general continuous outcome and specific survival outcomes have also been described in Hedges and Olkin (1985), Earle & Wells (2000), Parmar, Torri, & Stewart, (1998) and Srinivasan & Zhou (1993).

When individual studies used in a meta-analysis have very differing results, however, the results from systematic reviews may be less convincing and reliable. In an attempt to establish whether study results are consistent, reports on meta-analysis commonly present a statistical test of heterogeneity among studies used in a meta-analysis. This test seeks to determine whether there are genuine differences underlying the results of the studies, or whether the variation in these results is compatible with chance alone (i.e., homogeneity). A common statistical test used for this purpose is the Cochran's Chi-squared test or the *Q*-test (Whitehead & Whitehead, 1991; Cochran, 1954). It has been widely realized, however, that this test has poor power when the number of

studies in a meta-analysis is small, and excessive power to detect clinically insignificant heterogeneity when there are too many studies (Higgins & Thompson, 2002).

Addressing statistical heterogeneity of studies is one of the most fundamental aspects of many systematic reviews. The interpretative aspects of statistical inferences from a meta-analysis depend on the degree of heterogeneity of the studies used in the meta-analysis. Because the heterogeneity may determine the extent to which the conclusions of a meta-analysis can be generalized, it is important to quantify the extent of heterogeneity among a collection of studies. Realizing the potential limitations of a statistical test to characterize the degree of heterogeneity in a meta-analysis, Higgins and Thompson (2002) proposed a new measure of the extent of heterogeneity in a meta-analysis that overcomes the shortcomings of existing measures.

Their focus is on the impact of heterogeneity on the results of a meta-analysis and therefore, on the degree to which conclusion might be generalized to situations outside those investigated in the studies at hand. Their measure is easily interpretable by non-statisticians as the proportion of variation that was explained by the difference among studies. Further, the measure does not intrinsically depend on the number of studies or the type of outcome data, therefore offering the possibility that statistical heterogeneity can be compared across different meta-analyses with differing number of studies and types of outcome data.

In this article, several new indices are proposed that measure the heterogeneity from studies used in a meta-analysis. The proposed methodology can be regarded as a generalization of the index of heterogeneity proposed by Higgins and Thompson (2002). The difference among the proposed measures of heterogeneity are examined, along with the variation of each proposed measure when a large number of simulated meta-analyses are conducted. The proposed methodology is demonstrated by presenting an example to study possible cerebrospinal fluid (CSF) biomarkers that could be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal.

HETEROGENEITY IN META-ANALYSES

Indices of overall heterogeneity in a meta-analysis

Assume that a total of k studies are used in a meta-analysis to address a scientific question as represented by parameter θ . Let $\hat{\theta}_i$ be the estimate from the i -th study and $\hat{\sigma}_i^2$ be the associated estimated variance which is assumed to be known. Let $w_i = 1/\hat{\sigma}_i^2$ denote the precision of the estimate. In a classic fixed effect meta-analysis, θ_i 's are assumed identical and a summary estimate, $\hat{\theta}$, is computed to the common parameter as a weighted average of the study specific estimates, using the precisions as weights:

$$\hat{\theta} = \frac{\sum_{i=1}^k w_i \hat{\theta}_i}{\sum_{i=1}^k w_i}.$$

The variance of the summary estimate is given by

$$\hat{\sigma}_{\hat{\theta}}^2 = \frac{1}{\sum_{i=1}^k w_i}.$$

A random effects meta-analysis can be conceptualized by incorporating a random effect to account for the between-study variation, $N(0, \tau^2)$, into the estimated study-specific parameters, in addition to the within-study random variation, $N(0, \sigma_i^2)$. The summary estimate to the mean parameter across the distribution of the studies, $\hat{\theta}_r$, has exactly the same form as above, but with weights replaced by

$$\hat{w}_i^* = \frac{1}{(w_i^{-1} + \tau^2)}.$$

The estimated variance of the summary estimate is now given by

$$\hat{\sigma}_{\hat{\theta}_r}^2 = \frac{1}{\sum_{i=1}^k w_i^*}.$$

A test of homogeneity of the θ_i 's is given by

$$Q = \sum_{i=1}^k w_i (\hat{\theta}_i - \hat{\theta})^2,$$

which has a Chi-squared distribution with $k-1$ degrees of freedom under the assumption of homogeneity within the fixed effects model. Within the framework of the random effect model, a method of moment estimate to τ^2 can also be obtained as

$$\hat{\tau}^2 = \frac{Q - (k-1)}{\sum_{i=1}^k w_i - \frac{\sum_{i=1}^k w_i^2}{\sum_{i=1}^k w_i}}. \quad (1)$$

Higgins and Thompson (2002) proposed a simple index to quantify the overall heterogeneity among studies in a meta-analysis:

$$I^2 = \frac{\tau^2}{\tau^2 + \sigma^2},$$

where σ^2 is the shared within-study variance among individual studies, or when the studies have differing within-study variations, the typical within-study variance in the term of Higgins and Thompson (2002). This intuitive definition of the heterogeneity has several major advantages as compared to the standard statistical test based on Q . First, the definition of I^2 depends on the study specific estimates and is therefore based on the impact rather than the extent of heterogeneity in a meta-analysis. Second, the measure does not inherently depend on the number of studies in the meta-analysis. Third, the measure is not specific to a particular metric of treatment effect and therefore can be applied similarly irrespective of the type of outcome variables (e.g., dichotomous, continuous, and survival). Fourth, the measure is

easy to compute and has a very appealing interpretation as the percentage of the total variation across studies due to heterogeneity.

The estimation of overall heterogeneity among studies in a meta-analysis requires the estimate to both the between-study variation and the typical within-study variance. For the latter, Higgins and Thompson (2002) used the following estimator

$$\hat{\sigma}_{HT}^2 = \frac{(k-1) \sum_{i=1}^k w_i}{\left(\sum_{i=1}^k w_i \right)^2 - \sum_{i=1}^k w_i^2}.$$

This, along with the method of moment estimator $\hat{\tau}^2$, results in the index of overall heterogeneity

$$I_{HT}^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_{HT}^2} = \frac{Q - (k-1)}{Q} \quad (2)$$

Higgins and Thompson's (2002) intuitive conceptualization of the measure of heterogeneity is followed, and several new measures of heterogeneity are proposed. First, as pointed out by Takkouche et al. (Takkouche, Cadarso Surez, & Spiegelman, 1999), the typical within-study variance σ^2 can also be estimated by taking the reciprocal of the arithmetic mean weights:

$$\hat{\sigma}_T^2 = \frac{k}{\sum_{j=1}^k w_j}.$$

Combine this with the method of moment estimator $\hat{\tau}^2$ in Equation (1) to obtain another index of overall heterogeneity

$$I_T^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_T^2} = \frac{Q - (k-1)}{Q + 1 - \frac{k \sum_{j=1}^k w_j^2}{\left(\sum_{j=1}^k w_j \right)^2}}. \quad (3)$$

Another straightforward estimator to the typical within-study variance σ^2 can be obtained by simply averaging the within-study variances from all studies:

$$\hat{\sigma}_S^2 = \frac{\sum_{j=1}^k \frac{1}{w_j}}{k},$$

which, again along with the method of moment estimator $\hat{\tau}^2$ in Equation (1), results in another index of overall heterogeneity

$$\begin{aligned} I_S^2 &= \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_S^2} \\ &= \frac{Q - (k-1)}{Q - k + 1 + \sum_{j=1}^k \frac{1}{w_j} \left[\left(\sum_{j=1}^k w_j \right)^2 - \sum_{j=1}^k w_j^2 \right]}. \end{aligned} \quad (4)$$

These proposed indices of heterogeneity are set to 0 if $Q \leq (k-1)$. By Schwartz's inequality (Nobel & Daniel, 1977),

$$\left(\sum_{j=1}^k w_j \right)^2 \leq k \sum_{j=1}^k w_j^2$$

and

HETEROGENEITY IN META-ANALYSES

$$k^2 = \left(\sum_{j=1}^k \sqrt{w_j} \frac{1}{\sqrt{w_j}} \right)^2 \leq \sum_{j=1}^k w_j \sum_{j=1}^k \frac{1}{w_j}.$$

It then follows that

$$I_S^2 \leq I_T^2$$

and

$$I_{HT}^2 \leq I_T^2.$$

Notice that for all these indices of overall heterogeneity, although they have different denominators, they share the same numerator, which is $Q - (k - 1)$. If all within-study variations are exactly the same, then $I_{HT}^2 = I_T^2 = I_S^2$. Notice that the denominator of all these proposed overall measures of heterogeneity is the unconditional variance of the estimated effect from a typical study in the meta-analysis, which contains additive components due to the within-study variance (i.e., from between-patient variation within a study) and the between-study variation (i.e., heterogeneity).

Variation of the proposed overall measures of heterogeneity

Higgins and Thompson (2002) proposed several ways of estimating the variation associated with I_{HT}^2 . They recommended the use of an uncertainty interval based on the statistical significance of Q due to the appropriate nominal coverage in their simulation studies. Because the other measures of overall heterogeneity we proposed here also depend on Q , we use similar test-based methods (Miettinen, 1976) to estimate the variability associated with I_T^2 and I_S^2 as well. More specifically, let

$$Z = \sqrt{2Q} - \sqrt{2k - 3}.$$

Based on a well known normal approximation to Chi-squared distributions (Abramowitz & Stegun, 1965), when the degrees of freedom are large, Z follows approximately a standard

normal distribution. Therefore, if $\ln(Q)$ is assumed a normal distribution, by equating Z with $\frac{\ln(Q) - \ln(k - 1)}{SE(\ln(Q))}$, one can approximate the standard error of $\ln(Q)$ by

$$SE(\ln(Q)) = \frac{\ln(Q) - \ln(k - 1)}{\sqrt{2Q} - \sqrt{2k - 3}}.$$

This then results in a 95% uncertainty interval to I_T^2 as $[I_{T1}^2, I_{T2}^2]$, where

$$I_{T1}^2 = \frac{\exp(\ln(Q) - 1.96SE(\ln(Q))) - (k - 1)}{k \sum_{j=1}^k w_j^2 \exp(\ln(Q) - 1.96SE(\ln(Q))) + 1 - \left(\sum_{j=1}^k w_j \right)^2}$$

and

$$I_{T2}^2 = \frac{\exp(\ln(Q) + 1.96SE(\ln(Q))) - (k - 1)}{k \sum_{j=1}^k w_j^2 \exp(\ln(Q) + 1.96SE(\ln(Q))) + 1 - \left(\sum_{j=1}^k w_j \right)^2}$$

Similarly, a 95% uncertainty interval to I_S^2 is $[I_{S1}^2, I_{S2}^2]$, where

$$I_{S1}^2 = \frac{\exp(\ln(Q) - 1.96SE(\ln(Q))) - (k - 1)}{\exp(\ln(Q) - 1.96SE(\ln(Q))) - k + 1 + \sum_{j=1}^k \frac{1}{w_j} \left[\frac{\left(\sum_{j=1}^k w_j \right)^2 - \sum_{j=1}^k w_j^2}{k \sum_{j=1}^k w_j} \right]}$$

and

$$I_{S2}^2 = \frac{\exp(\ln(Q) + 1.96SE(\ln(Q))) - (k - 1)}{\exp(\ln(Q) + 1.96SE(\ln(Q))) - k + 1 + \sum_{j=1}^k \frac{1}{w_j} \left[\frac{\left(\sum_{j=1}^k w_j \right)^2 - \sum_{j=1}^k w_j^2}{k \sum_{j=1}^k w_j} \right]}.$$

Comparison of the proposed overall indices of heterogeneity

Although mathematically, $I_S^2 \leq I_T^2$, $I_{HT}^2 \leq I_T^2$, it is important to understand how different these measures are when they are used to measure the overall heterogeneity in a meta-analysis and how much variation each index has when a large number of meta-analyses are conducted. Given the fact that when all studies have exactly the same degree of within-study variation, i.e., when all w_i 's are the same, these measures are identical to each other, we anticipate that these measures will be close to each other when the difference among within-study variations is relatively small.

A simulation study is performed to look at the performance of our proposed measures of overall and study-specific heterogeneity. We first examined the distributions and consistency of three different measures of overall heterogeneity, I_S^2 , I_T^2 , and I_{HT}^2 , over a large number of simulated meta-analyses. Assume that the between-study variance is $\tau^2 = 4$. The number of studies in each simulated meta-

analysis is $k = 3s + 1$ for $s = 2, 4, 8$, and 12. In each simulated meta-analysis, three different within-study variances are assumed such that the precision w_i is either $0.5 + \nu$, or $0.5 + 2\nu$, or $0.5 + 3\nu$ for a range of ν . More specifically, among $k = 3s + 1$ studies in the meta-analysis, $s + 1$ studies have within-study precision $0.5 + \nu$, and the other $2s$ studies are equally distributed with study precision $0.5 + 2\nu$ and $0.5 + 3\nu$. A random effect model was used to generate the study-specific estimates such that the between-study component was generated from the normal distribution $N(5, 4)$ through a linear transformation of the SAS Institute function RANNOR (SAS, 1999). One thousand independent simulated meta-analyses were performed such that study specific estimates from each individual simulation were independently generated. Table 1 presents the mean and the associated standard error for the three proposed measures of overall heterogeneity over 1000 simulated meta-analyses as a function of k and ν (notice that parameter ν here indicates a measure of difference among the study precisions).

HETEROGENEITY IN META-ANALYSES

From the simulated meta-analyses, it is clear that three different measures of overall heterogeneity are very consistent. In fact, under the assumption that the three measures of heterogeneity are estimating the same underlying trait of heterogeneity, the intraclass correlation coefficient (ICC) (Shrout & Fleiss, 1979) was computed over 1,000 simulated meta-analyses for each choice of k and v . All these computed ICCs were at least 0.99, indicating extremely high consistency among these measures.

Application to biomarker studies in Alzheimer's disease

An application to the proposed overall measures of heterogeneity is presented to study possible biomarkers that can be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal. Researchers in Alzheimer's disease have identified Apolipoprotein E4 (ApoE4) alleles as a crucial genetic risk factor of AD (Myers, Schaefer, Wilson, et al., 1996). Although the pathological hallmarks of AD are the neurofibrillary tangles and the senile plaques in the brain (Braak & Braak, 1991, McKell, Price, Miller, Grant, Xiong, Berg, & Morris, 2004), the diagnosis of AD in living patients is still largely a clinical judgment based on careful neurological and/or neuropsychological examinations combined with results from other clinical tests.

Therefore, the search for biomarkers that could be used to diagnose AD from normal aging has been one of the primary research activities in AD. In several publications (Fagan, Roe, Xiong, et al., 2007, Sunderland, Linker, Mirza, et al., 2003), subjects with AD have been found to have decreased level of cerebrospinal fluid (CSF) β -amyloid₄₂ as compared to subjects with normal aging. Because AD is a progressive neurodegenerative disorder that leads to the death of brain cells that cannot be replaced once lost, it is important to assess the potential of these biomarkers to identify subjects that are at high risk of AD while they are still cognitively normal. The importance of such biomarkers is further highlighted by the fact that no pharmaceutical treatments are effective for

the disease's later stages. Thus, whether CSF β -amyloid₄₂ is decreased among subjects of normal aging who are ApoE4 positive as compared to these who are ApoE4 negative is studied.

Although many publications have compared CSF β -amyloid₄₂ level between subjects with AD and these with normal aging (Fagan, Roe, Xiong, et al., 2007, Sunderland, Linker, Mirza et al., 2003), very few have actually reported CSF β -amyloid₄₂ as a function of ApoE4 status among subjects who were still cognitively normal. As a matter of fact, our comprehensive MEDLINE search identified a total of 6 published studies on CSF β -amyloid₄₂ during the period of 1990 to 2007 which actually reported summary statistics as a function of ApoE4 status for subjects who were not demented (Prince, Zetterberg, Andreasen, et al. 2004, Sunderland, Mirza, Putnam, et al., 2004, Jensen, Schroder, Blomberg et al., 1999, Andreasen, Hesse, Davidson et al., 1999, Tapiola, Pirttila, Mehta, et al., 2000, Riemenschneider, Schmolke, Lautenschlager, et al, 2000). The summary statistics reported from these six published studies are presented in Table 2 (summary statistics from study by Prince, Zetterberg, Andreasen, et al., 2004) was obtained through eye-balling because only a graphical presentation on summary statistics was available in the publication).

Based on the proposed methodology and a random effect model, the pooled estimate to the mean difference of CSF β -amyloid₄₂ between subjects of normal aging who are ApoE4 positive and subjects who are ApoE4 negative is -31.69 pg/mL, and an asymptotic 95% confidence interval estimate to the mean difference of CSF β -amyloid₄₂ is from -128.93 pg/mL to 65.56 pg/mL. The observed significance level for the observed mean difference is 0.407. The measures of overall heterogeneity from this meta-analysis are estimated as $I_{HT}^2 = 0.56$, $I_T^2 = 0.66$, and $I_S^2 = 0.20$, respectively, indicating from low to moderate degree of heterogeneity among studies used in the meta-analysis (Higgins, Thompson, Deeks et al., 2003). Further, an estimated 95%

Table 1. Three Measures of Overall Heterogeneity from 1000 Simulated Meta-analyses
 (k = the number of studies in meta-analyses,
 $(0.5+\nu, 0.5+2\nu, 0.5+3\nu)$ = the three within-study precisions)

k	ν	Mean I_{HT}^2 (SE) (%)	Mean I_T^2 (SE) (%)	Mean I_S^2 (SE) (%)
7	0	39.29 (0.85)	39.29 (0.85)	39.29 (0.85)
7	0.5	73.83 (0.58)	74.06 (0.57)	72.68 (0.58)
7	1.0	83.73 (0.45)	83.97 (0.45)	82.46 (0.46)
7	1.5	88.48 (0.31)	88.70 (0.30)	87.28 (0.32)
7	2.0	90.17 (0.29)	90.38 (0.29)	89.04 (0.31)
13	0	43.04 (0.66)	43.04 (0.66)	43.04 (0.66)
13	0.5	79.33 (0.34)	79.43 (0.34)	78.14 (0.35)
13	1.0	87.27 (0.19)	87.37 (0.19)	85.99 (0.21)
13	1.5	90.80 (0.15)	90.89 (0.15)	89.65 (0.16)
13	2.0	92.56 (0.14)	92.64 (0.14)	91.53 (0.15)
25	0.0	46.16 (0.50)	46.16 (0.50)	46.16 (0.50)
25	0.5	81.25 (0.18)	81.30 (0.18)	80.05 (0.19)
25	1.0	88.62 (0.13)	88.66 (0.12)	87.40 (0.14)
25	1.5	91.89 90.08)	91.93 (0.08)	90.81 (0.09)
25	2.0	93.69 (0.07)	93.73 (0.07)	92.76 (0.08)
37	0.0	47.41 (0.39)	47.41 (0.39)	47.41 (0.39)
37	0.5	81.88 (0.15)	91.91 (0.15)	80.69 (0.16)
37	1.0	89.22 (0.09)	89.25 (0.09)	88.03 (0.10)
37	1.5	92.16 (0.07)	92.18 (0.07)	91.10 (0.08)
37	2.0	93.94 (0.05)	93.96 (0.05)	93.03 (0.06)

HETEROGENEITY IN META-ANALYSES

Table 2. Reported Summary Statistics from Six Studies on CSF β -amyloid₄₂ (in pg/mL) as a Function of ApoE4 Genotype

(Author = the first author of the study, Year = the year of the publication,
 n_+ = the sample size of subjects who are ApoE4 positive,
 n_- = the sample size of subjects who are ApoE4 negative,
Mean (SD)+ = mean (standard deviation) in subjects who are ApoE4 positive,
Mean (SD)- = mean (standard deviation) in subjects who are ApoE4 negative)

Author	Year	n_+ / n_-	Mean (SD)+	Mean (SD)-
Andreasen N	1999	8/13	1641.00 (587.00)	1702.00 (339.00)
Jensen M	1999	4/20	365.72 (85.79)	329.60 (139.97)
Tapiola T	2000	13/25	500.00 (211.00)	522.00 (136.00)
Riemenschneide M	2000	3/15	914.67 (11.37)	860.00 (194.00)
Sunderland T	2004	57/85	389.00 (108.00)	443.00 (109.00)
Prince JA	2004	32/86	697.00 (228.00)	840.00 (185.00)

uncertainty interval for I_{HT}^2 is from 0.00 to 0.82, an estimated 95% uncertainty interval for I_T^2 is from 0.00 to 0.88, and an estimated uncertainty interval for I_S^2 is from 0.00 to 0.48 (the uncertainty intervals were truncated to 0 when the left limits were negative as similarly recommended in Higgins & Thompson, 2002).

If the heterogeneity is ignored in the meta-analysis, i.e., the between-study variance τ^2 is treated as 0 (therefore $I_{HT}^2 = I_T^2 = I_S^2 = 0$), then a fixed effect model would be used for the meta-analysis. The estimated overall mean difference of CSF β -amyloid₄₂ between subjects of normal aging who are ApoE4 positive and subjects who are ApoE4 negative under the fixed effect model is -45.35 pg/mL. An asymptotic 95% confidence interval estimate to the mean difference of CSF β -amyloid₄₂ under the fixed effect model is from -74.89 pg/mL to -15.82 pg/mL, suggesting a statistically significant difference at a 5% significance level on CSF β -amyloid₄₂ between subjects of normal aging who are ApoE4 positive and subjects who are ApoE4 negative. This discrepancy on the statistical inference between the two approaches that either take into

account of heterogeneity (i.e., random effect models) or ignore the heterogeneity (i.e., fixed effect models) further highlights the importance to assess the heterogeneity in meta-analyses.

Conclusion

We proposed several new indices that measure the overall heterogeneity among studies used in a meta-analysis. By estimating the typical within-study precisions, we developed indices that measure the degree of heterogeneity among studies by their impact to the overall conclusion of the meta-analysis. The proposed methodology can be regarded as a generalization of the index of heterogeneity proposed by Higgins and Thompson (2004). We assessed the variation associated with each proposed index of heterogeneity through a large simulation study of 1000 meta-analyses for a range of relevant parameters. We also examined the difference among the proposed overall measures of heterogeneity when a large number of meta-analyses were conducted. We found that these different indices provided highly consistent results in measuring the overall heterogeneity in meta-analyses. Finally, we demonstrated our proposed methodology by presenting an example to study possible biomarkers that could

be used to identify subjects with high risk of developing Alzheimer's disease (AD) when they are still cognitively normal. The inconsistent statistical inferences to this real world example based on statistical approaches with or without taking into account of heterogeneity highlight the crucial role heterogeneity plays in meta-analyses.

References

- Abramowitz, M., & Stegun, I. A. (1965). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. NY: Dover.
- Andreasen, N., Hesse, C., Davidsson, P., et al. (1999). Cerebrospinal fluid β -amyloid₍₁₋₄₂₎ in Alzheimer's disease: differences between early- and late-onset Alzheimer's disease and stability during the course of disease. *Archives in Neurology*, 56, 673-680.
- Antman, E. M., Lau, J., Kupelnick, B., Mosteller, F., & Chalmers, T. C. (1992). A comparison of results of meta-analyses of randomized control trials and recommendations of clinical experts: treatment for myocardial infarction. *Journal of the American Medical Association*, 268, 240-248.
- Braak, H., & Braak, E. (1991). Neuropathologic staging of Alzheimer-related changes. *Acta Neuropathologica*, 82, 239-259.
- Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, 10, 101-129.
- DerSimonian, R., Laird, N. M. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7, 177-188.
- Earle, C. C., & Wells, G. A. (2000). An assessment of methods to combine published survival curves. *Medical Decision Making*, 20, 104-111.
- Egger, M., & Davey Smith, G. (1997). Meta-analysis: potentials and promise. *British Medical Journal*, 315, 1371-1374.
- Egger, M., Davey Smith, G., & Phillips, A. N. (1997). Meta-analysis; principles and procedures. *British Medical Journal*, 315, 1533-1537.
- Fagan, A. M., Roe, C. M., Xiong, C., et al. (2007). Cerebrospinal fluid tau/ β -amyloid₄₂ ratio as a prediction of cognitive decline in nondemented older adults. *Archives in Neurology*, 64. (doi:10.1001/archneur.64.3.noc60123).
- Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, 17, 841-856.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando: Academic Press.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1558.
- Higgins, J. P. T., Thompson, S. G., Deeks, J. J., et al. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal*, 327, 557-560.
- Jensen, M., Schroder, J., Blomberg, M., et al. (1999). Cerebrospinal fluid A β 42 is increased early in sporadic Alzheimer's disease and declines with disease progression. *Ann Neurol*, 45, 504-511.
- Laird, N. M., & Mosteller, F. (1999). Some statistical methods for combining experimental results. *The International Journal of Technology Assessment in Health Care*, 6, 5-30.
- Mantel, N., Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719-748.
- McKeel, D. W., Jr., Price, J. L., Miller, J. P., Grant, E. A., Xiong, C., Berg, L., & Morris, J. C. (2004). Neuropathologic criteria for diagnosing Alzheimer disease in persons with pure dementia of Alzheimer type. *Journal of Neuropathology and Experimental Neuropathology*, 63(10), 1028-1037.
- Miettinen, O. (1976) Estimability and estimation in case-referent studies. *American Journal of Epidemiology*, 103, 226-235.
- Myers, R. H., Schaefer, E. J., Wilson, P. W. F., et al. (1996). Apolipoprotein E ϵ 4 association with dementia in a population-based study: The Framingham study. *Neurology*, 46, 673-677.

HETEROGENEITY IN META-ANALYSES

Noble, B., & Daniel, J. W. (1977). *Applied linear algebra*. Englewood Cliffs, NJ: Prentice-Hall.

Parmar, M. K. B., Torri, V., & Stewart, L. A. (1998). Extracting summary statistics to perform meta-analyses of published literature for survival endpoints. *Statistics in Medicine*, 17, 2815-2834.

Prince, J. A., Zetterberg, H., Andreasen, N., et al. (2004). APOE ϵ 4 allele is associated with reduced cerebrospinal fluid levels of A β 42. *Neurology*, 62, 2116-2118.

Riemenschneider, M., Schmolke, M., Lautenschlager, N., et al. (2000). Cerebrospinal beta-amyloid₍₁₋₄₂₎ in early Alzheimer's disease: association with apolipoprotein E genotype and cognitive decline. *Neuroscience Letters*, 284, 85-88.

Sackett, D. L., & Haynes, R. B. (1995). On the need for evidence-based medicine. *Evidence-Based Medicine*, 1, 5-6. SAS Institute, Inc. (199).

SAS language (Ver. 8). Cary, NC

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.

Srinivasan, C., & Zhou, M. (1993). A note on pooling Kaplan-Meier estimators. *Biometrics*, 49, 861-864.

Sunderland, T., Linker, G., Mirza, N., et al. (2003). Decreased β -amyloid₁₋₄₂ and increased tau levels in cerebrospinal fluid of patients with Alzheimer's disease. *JAMA*, 289, 2094-2103.

Sunderland, T., Mirza, N., Putnam, K. T., et al. (2004). Cerebrospinal fluid β -amyloid₁₋₄₂ and tau in control subjects at risk for Alzheimer's disease: The effect of apoE ϵ 4 allele. *Biological Psychiatry*, 56, 670-676.

Takkouche, B., Cadarso-Surez, C., Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*, 150, 206-215.

Tapiola, T., Pirttila, T., Mehta, P. D., et al. (2000). Relationship between apoE genotype and CSF β -amyloid(1-42) and tau in patients with probable and definite Alzheimer's disease. *Neurobiology of Aging* 2000, 21, 735-740.

Whitehead A., & Whitehead, J. (1991). A general parametric approach to the meta-analysis of clinical trials. *Statistics in Medicine*, 10, 1665-1677.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1985). Beta blockade during and after myocardial infarction: an overview of the randomized trials. *Progress in Cardiovascular Diseases*. 27, 335-371.

When Sensitivity is a Function of Age and Time Spent in the Preclinical State in Periodic Cancer Screening

Dongfeng Wu
University of Louisville

Ricolindo L. Cariño Xiaoqin Wu
Mississippi State University

Probability models are extended for periodic cancer screening trials to model sensitivity when it is changing with an individual's age and time spent in the preclinical state. Wu et al. (2005) showed that sensitivity is monotone increasing with age, but intuitively, sensitivity is also a function of the time one has spent in the preclinical stage. This allows us to infer sensitivity at a late stage, just before symptoms manifest. We developed the probability model and applied Bayesian inference to the HIP study group data. The methodology we developed is also applicable to other kinds of chronic diseases.

Keywords: Periodic screening exam, breast cancer, sensitivity, sojourn time, transition probability, incidence.

Introduction

Early detection and treatment may be effective ways to decrease mortality rate from cancer. The primary technique for early detection is screening exams. According to a recent report of the National Institute of Health (NIH 2000), breast cancer is the most common form of cancer among women in the United States and the second leading cause of cancer deaths among women. In the past four decades, seven major randomized controlled breast cancer screening trials have been carried out in North America and Europe.

In a screening program, a large group of asymptomatic individuals are enrolled in the program to detect the presence of a specific disease. The natural history of the disease for an individual is assumed to follow a progressive stochastic model, which consists of three states,

denoted by $S_0 \rightarrow S_p \rightarrow S_c$, corresponding, respectively, to the disease-free state; the preclinical disease state, in which an asymptomatic individual unknowingly has disease that the screening exam can detect; and the clinical state when the disease manifests itself in clinical symptoms. The *sensitivity* is the probability that the screening exam is positive given that the individual is in the preclinical stage.

The *sojourn time* refers to the time beginning when the disease first develops until the manifestation of clinical symptoms, that is $(S_c - S_p)$. The *transition probability* into the preclinical stage is the probability density function of making transition from the disease-free to the preclinical state. Knowledge of the sensitivity of the screening modality is necessary for evaluating the predictive performance of a screening exam. The screening sensitivity may depend on a variety of factors, including age, position, location and size of the tumor, the experience of the radiologist, etc. For example, recent studies indicate that the sensitivity of mammography increases with age at diagnosis (Wu et al., 2005; Shapiro et al., 1988; Miller et al., 1992a, 1992b), attributable to the fact that breast tissue tends to be more dense and fibrous in younger women, and more soft and fatty in older women (Kerlikowske, et al., 1996).

Dongfeng Wu is an Associate Professor in the Department of Bioinformatics and Biostatistics. She is a member of the Editorial Board of *JMASM*. E-mail: dongfeng.wu@louisville.edu. Ricolindo L. Carino is on the faculty at the Center for Advanced Vehicular Systems. Xiaoqin Wu is an Assistant Professor in the Department of Mathematics, Computer and Information Sciences.

In Walter and Day (1983), it was found that sensitivity is negatively correlated with the sojourn time. Intuitively, when the tumor cell is just formed, the sensitivity is very small, while at the late stage, that is, the preclinical stage comes to an end and the clinical stage will start soon, the sensitivity is close to one. In Wu, et al. (2005), the sensitivity was modeled as a function of age, while transition probability is age-dependent as well. Previous result are extended by investigating changes in the sensitivity from simultaneous variation of age and time spent in the preclinical stage.

The Model

Consider a cohort of initially asymptomatic individuals who enroll in a screening program. The sensitivity is denoted by $\beta(t, s|T)$, where t is the individual's age at the screening exam, s is the time duration that one has been stayed in the preclinical state, and T represents the sojourn time in the preclinical state. Define $w(t)dt$ as the probability of a transition from S_0 to S_p during $(t, t+dt)$. Let $q(t)$ be the probability density function of the sojourn time in S_p . Let $Q(z) = \int_z^\infty q(x)dx$, that is, $Q(z)$ is the survivor function of the sojourn time in the preclinical state S_p . Throughout this paper, the time variable t represents the participating individual's age. If random variables T and S are the duration times in S_0 and S_p respectively, then an individual will enter the clinical state S_c at age $T+S$, the probability density function of $(T+S)$ is

$$I(t) = \int_0^t w(x)q(t-x)dx.$$

$I(t)$ is the observable incidence of clinical cases if no intervention exists.

Consider a cohort of women in the study group who are all aged t_0 at study entry, and a protocol for K ordered screening examinations occurring at ages $t_0 < t_1 < \dots < t_{K-1}$, where $t_i = t_0 + i$ for annual screening exams. Define the i -th screening interval as the time interval between the i -th and the $(i+1)$ -th screening exams (t_{i-1}, t_i) , $i=1, 2, \dots, K-1$. The i -th generation of individuals consists of those who

enter S_p during this interval. The 0-th generation includes all who enter S_p before the initial screening exam, and we let $t_{-1} \equiv 0$.

For each screening exam, let n_{i,t_0} be the total number of individuals in this cohort examined at the i -th screening, s_{i,t_0} is the number of cases detected at the i -th screening exam, and r_{i,t_0} is the number of cases diagnosed in the clinical state S_c within the interval (t_{i-1}, t_i) . The latter cases are called interval cases.

Let D_{k,t_0} be the probability that an individual will be diagnosed at the k -th scheduled exam (at which her age is $t_{k-1} = t_0 + k - 1$) given that she is already in the preclinical state. The probability that an individual in S_p is detected at the first scheduled exam (i.e. $k = 1$) at age t_0 is

$$D_{1,t_0} = \int_0^{t_0} w(x) \int_{t_0-x}^\infty q(t) \beta(t_0, t_0 - x | t) dt dx. \quad (1)$$

The double integral in equation (1) arises because she must have entered the preclinical state S_p before t_0 , remained in that state at least until t_0 , and with the sensitivity changing with the sojourn time and the time spent in the preclinical stage, as well as with age.

Consider an i -th generation individual who was diagnosed at the k -th screening exam ($1 \leq i < k$). There are two possibilities: one is that she passed her previous $(k-i-1)$ exams undetected and had a sojourn time of at least $(t_{k-1} - x)$, where $x \in (t_{i-1}, t_i)$ is her onset of the preclinical state S_p ; the other possibility is that she entered S_p in the $(k-1)$ -th screening interval (t_{k-2}, t_{k-1}) . Hence, the probability is

$$\begin{aligned}
 D_{k,t_0} = & \sum_{i=0}^{k-2} \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{\infty} q(t) \left(\prod_{j=i}^{k-2} [1 - \beta(t_j, t_j - x | t)] \right) \beta(t_{k-1}, t_{k-1} - x | t) dt dx \\
 & + \int_{t_{k-2}}^{t_{k-1}} w(x) \int_{t_{k-1}-x}^{\infty} q(t) \beta(t_{k-1}, t_{k-1} - x | t) dt dx, \\
 & \text{for } \forall k = 2, \dots, K.
 \end{aligned} \tag{2}$$

Let $I_{k,t_0}(t)dt$ be the probability that an individual enters the state S_c at a small age interval $(t, t + dt)$, where $t \in (t_{k-1}, t_k)$. If this woman was in the i -th generation, $i < k$, then she must have gone through her previous $(k-i)$ screening exams without being detected, and had a sojourn time $(t-x)$, where x is her onset age of the preclinical state S_p . Another possibility is that she may have entered S_p after the k -th exam and incident at age t . Hence for $\forall t \in (t_{k-1}, t_k)$,

$$\begin{aligned}
 I_{k,t_0}(t) = & \sum_{i=0}^{k-1} \int_{t_{i-1}}^{t_i} w(x) q(t-x) \left(\prod_{j=i}^{k-1} [1 - \beta(t_j, t_j - x | t-x)] \right) dx \\
 & + \int_{t_k}^t w(x) q(t-x) dx.
 \end{aligned} \tag{3}$$

Therefore, for any $k = 1, \dots, K$, the probability of being incident in the k -th screening interval (t_{k-1}, t_k) is

$$\begin{aligned}
 I_{k,t_0} = & \int_{t_{k-1}}^{t_k} I_{k,t_0}(t) dt = \\
 & \sum_{i=0}^{k-1} \int_{t_{i-1}}^{t_i} w(x) \int_{t_{k-1}-x}^{t_k-x} q(t) \left(\prod_{j=i}^{k-1} [1 - \beta(t_j, t_j - x | t)] \right) dt dx \\
 & + \int_{t_{k-1}}^{t_k} w(x) [1 - Q(t_k - x)] dx.
 \end{aligned} \tag{4}$$

The likelihood function for this cohort of women is

$$\begin{aligned}
 L(\cdot | t_0) = & \prod_{k=1}^K D_{k,t_0}^{S_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1 - D_{k,t_0}^{S_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0} - S_{k,t_0} - r_{k,t_0}}
 \end{aligned} \tag{5}$$

The full likelihood for the study group across all ages is

$$L = \prod_{t_0} \prod_{k=1}^K D_{k,t_0}^{S_{k,t_0}} I_{k,t_0}^{r_{k,t_0}} (1 - D_{k,t_0}^{S_{k,t_0}} - I_{k,t_0}^{r_{k,t_0}})^{n_{k,t_0} - S_{k,t_0} - r_{k,t_0}} \tag{6}$$

The age effect and the time spent in the preclinical state were modeled in the sensitivity in the following way. The sensitivity is associated with age by a logistic link, and it is associated with the time spent in the preclinical state by a linear function. Let

$$\beta(t, s | T) = \frac{1}{1 + \exp(-b_0 - b_1 * (t - \bar{t}))} \times \frac{s}{T}, \tag{7}$$

where t is an individual's age at the screening exam, s is the time one has spent in the preclinical state, T is the sojourn time in the preclinical state, and $\bar{t}$ is the average age at entry in the whole study group. If $b_1 > 0$, $\beta(t)$ will be a monotone increasing function of age t .

The transition probability density function $w(t)$ is the instantaneous probability of a transition from S_0 to S_p . The integral $\int_0^{\infty} w(t) dt$ represents a lifetime risk for a healthy female to transit into the preclinical state. According to the NCI's SEER database (Ries et al. 2002), a woman's lifetime risk of being diagnosed with breast cancer is 15.7%, which is less than a woman's lifetime risk of entering the preclinical disease state.

PERIODIC CANCER SCREENING

Table 1. Bayesian posterior estimates for the HIP data

Parameters	Median	Mean	S.E.
b_0	1.730	1.708	0.984
b_1	0.084	0.083	0.072
μ	4.384	4.392	0.065
σ^2	0.235	0.253	0.095
κ	1.744	2.126	1.004
ρ	0.381	0.366	0.104

Table 2. Bayesian posterior estimates for the sensitivity β at the end of the preclinical state

Age	Median	Mean	S.E.
40	0.644	0.628	0.215
45	0.736	0.716	0.178
50	0.829	0.789	0.150
55	0.886	0.841	0.129
60	0.917	0.877	0.115
65	0.940	0.899	0.109

Table 3. Bayesian posterior estimates for the transition probability aw .

Age	Median	Mean	S.E.
40	1.400	1.407	0.243
45	1.734	1.738	0.210
50	1.994	2.006	0.254
55	2.171	2.193	0.335
60	2.267	2.296	0.404
65	2.286	2.322	0.441

^a The unit is 10^{-3} .

Hence, 20% is chosen as a reasonable upper bound.

$$w(t) = \frac{0.2}{\sigma t \sqrt{2\pi}} \exp\left\{-\frac{(\log t - \mu)^2}{2\sigma^2}\right\} \quad (8)$$

which is the PDF of lognormal(μ , σ^2) multiplied by 20%. Here, $w(t)$ is a sub-density function, where μ and σ^2 are parameters to be estimated.

The log logistic distribution is adopted to model the sojourn time in the preclinical state,

$$q(x) = \frac{\kappa x^{\kappa-1} \rho^\kappa}{[1 + (x\rho)^\kappa]^2}, x > 0, \quad (9)$$

where x is the sojourn time, and κ and ρ are positive parameters representing the scale and location in the log logistic family. An advantage of this family over the exponential is that it has two parameters and is more robust in the tails.

Figure 1: Density Curve for the posterior samples.

Figure 2: Posterior quantiles (5%, 50%, and 95%) of sensitivity and transition probability.

Another advantage of this family is that its relatively simple form is achieved for the survivor function and the hazard function. Its first moment can be calculated directly from

$$E(X) = \frac{\pi}{\rho\kappa} \csc\left(\frac{\pi}{\kappa}\right).$$

For the r -th moment to exist, $\kappa > r$ is required. For justifications on how these age effect functions are chosen, see Wu, et al. (2005).

Simulation Procedure and Results

The analysis of the HIP study data is now described based on the likelihood function and probability calculation derived above. In the proposed model, there are six unknown parameters, that is, $\theta = (b_0, b_1, \mu, \sigma^2, \kappa, \rho)$. Theoretically, the parameters have a domain of either $(-\infty, \infty)$ or $(0, \infty)$. The practical meaning of these parameters will limit them to a finite range. The range for each of them was identified as: $0 < b_0 < 5$, $-0.2 < b_1 < 0.2$, $3.5 < \mu < 4.5$, $0 < \sigma^2 < 1$, $0.1 < \rho < 2.0$, and $1 < \kappa < 5$. For justifications of these ranges, see Wu, et al. (2005).

Markov Chain Monte Carlo (MCMC) was used to generate a random sample from the joint posterior distribution of the parameters for Bayesian inference. The posterior simulation was partitioned into four sub-chains, sampling the posterior for (b_0, b_1) , μ, σ^2 , (κ, ρ) separately.

A noninformative bivariate normal prior for (b_0, b_1) was chosen, that is, a bivariate normal distribution with mean vector $(0, 0)$ and variance equal to 10^{10} times the identity matrix. A noninformative normal prior was chosen for μ , namely $N(0, 10^{10})$. The prior for σ^2 was uniform $(0, 1)$, and the prior distributions for κ and ρ were uniform $(1, 5)$ and uniform $(0, 2)$, respectively. The two-dimensional integrals in the likelihood function do not have an analytical form. The trapezoidal rule was used to evaluate them when calculating the likelihood.

The MCMC was run for 4,800 steps, with a burn-in of 3,200 iterations. After the burn-in time, the posteriors were sampled every 20 steps, giving 80 posterior samples for the

parameter vector θ . Sixteen chains were simulated, each with different starting values that are over dispersed with respect to the target distribution. The two-dimensional integrals in the likelihood function are very time consuming. The chains were simulated in parallel on a Linux cluster, taking 192 hours to complete. The 80 posterior samples from each of the 16 chains were pooled for the analysis, giving a total of 1280 posterior samples.

The Bayesian posterior estimates of θ for the HIP study data are shown in Table 1. Sensitivity at the end of the preclinical state appears to increase with age. If $s \rightarrow T$, that is, if the time spent in the preclinical state s converges to the sojourn time T , then the sensitivity in the late stage of the preclinical state can be estimated. This trend is obvious from the quantile plot of the saved posterior samples of the parameters in Figure 1. In the HIP data, the posterior mean sensitivity increases from 0.628 to 0.899 from ages 40 to 65 years, and the posterior standard error drops from 0.215 to 0.109. In fact, the posterior error of sensitivity was monotone decreasing as age increases.

The age-dependent transition probability is itself a sub-pdf from our model construction. The posterior density curve of the transition probability could be seen from Figure 1. The posterior mean of the transition probability varies from 1.407×10^{-3} to 2.322×10^{-3} . The transition probability is not a monotone function of age; it has a single maximum at age 64. If the posterior means for the parameters κ and ρ were used, then the posterior mean sojourn time is 4.06 years, with a mode of 1.70 years.

Discussion

The previous model in Wu, et al. (2005) is extended by adding one element in the sensitivity function, namely, the time spent in the preclinical state, with the sojourn time. Walter and Day (1983) found that the sensitivity and the sojourn time distribution were negatively correlated for a given incidence. Our modeling should be able to handle that problem, as our sensitivity was defined as a function of the sojourn time as well, and it was reciprocal to the sojourn time. More complicated models

could be explored, but the computation time will be greatly increased, and simulation would take too long to be finished.

The result was compared with Wu, et al. (2005). It was found that the late stage sensitivity was slightly higher, with a slightly smaller standard error. It was also found that the transition probability was changing less across different ages, with a slightly smaller standard error. However, the mean sojourn time is much longer than the previous result.

Acknowledgements

This research was partially supported by the National Institutes of Health grant CA-115012.

References

- Baker, S. G. (1998). Evaluating the age to begin periodic breast cancer screening using data from a few regularly scheduled screenings. *Biometrics*, 54, 1569-1578.
- Chen, T. H. H., Kuo, H. S., Yen, M. F., Lai, M. S., Tabar, L., & Duffy, S. W. (2000). Estimation of sojourn time in chronic disease screening without data on interval cases. *Biometrics*, 56, 167-172.
- Cox, D. R., & Oakes, D. (1984). *Analysis of survival data*. NY: Chapman & Hall/CRC.
- Day, N. E., & Walter, S. (1984). Simplified models of screening for chronic disease: estimation procedures from mass screening programmes. *Biometrics*, 40, 1-13.
- Eddy, D. M. (1980). *Screening for cancer: Theory, analysis and design*. Englewood Cliffs, NJ: Prentice Hall.
- Kerlikowske, K., Grady, D., Barclay, J., Sickles, E. A., & Ernster, V. (1996). Effect of age, breast density, and family history on the sensitivity of first screening mammography. *Journal of the American Medical Association*, 276, 33-38.
- Lee, S. J. & Zelen, M. (1998). Scheduling periodic examinations for the early detection of disease: Applications to breast cancer. *Journal of the American Statistical Association*, 93, 1271-1281.
- Miller, A. B., Baines C. J., To, T. & Wall, C. (1992a). Canadian national breast screening study: 1. Breast cancer detection and death rates among women aged 40 to 49 years. *Canadian Medical Association Journal*, 147(10), 1459-76.
- Miller, A. B., Baines, C.J., To, T. & Wall, C. (1992b). Canadian National Breast Screening Study: 2. Breast cancer detection and death rates among women aged 50 to 59 years. *Canadian Medical Association Journal*, 147(10), 1477-88.
- National Institute of Health (2000). *NIH Publication No. 00-1556*, 12/12/2000.
- Shapiro, S., Venet, W., Strax, P., & Venet L. (1988) *Periodic screening for breast Cancer. The health insurance plan project and its sequelae, 1963-1986*. Baltimore: The Johns Hopkins University Press.
- Shen, Y., Wu, D., & Zelen, M. (2001). Testing the independence of two diagnostic tests. *Biometrics*, 57, 1009-1017.
- Straatman, H., Peer, P. G. M. & Verbeek, A. L. M. (1997). Estimating lead time and sensitivity in a screening program without estimating the incidence in the screened group. *Biometrics*, 53, 217-229.
- Walter, S. D. & Day, N. E. (1983). Estimation of the duration of a preclinical disease state using screening data. *American Journal of Epidemiology*, 118, 856-86.
- Wu, D., Rosner, G. L., and Broemeling, L. D. (2005). MLE and Bayesian inference of age-dependent sensitivity and transition probability in periodic screening. *Biometrics*. Vol.61, No.4, 1056-1063.
- Zelen, M. (1993). Optimal scheduling of examinations for the early detection of disease. *Biometrika*, 80, 279-93.

Log-linear Model to Assess Socioeconomic and Environmental Factors with Childhood Diarrhea Using Hospital Based Surveillance

Krishnan Rajendran Thandavarayan Ramamurthy Sujit Kumar Bhattacharya
National Institute of Cholera and Enteric Diseases P-33

Categorical outcomes with environment factors analyzed by log linear model are frequent in the environmental epidemiological literature. Epidemiological and socio-economical factors were obtained on 1,119 children below the age of 5 from Infectious Diseases Hospital (IDH) at the Kolkata, India. Significant associations of diarrhea were observed in the rural areas with family income, father's occupation as a daily labor, literacy of parents, non-cemented floor and wall constructed of mud, and type of storage (wide mouthed earthen pot). The results of the study with specific Log linear model confirm environmental factors were important implications for childhood diarrhea in the rural community. To reduce the diseases burden, the intervention strategies such as education, improvement of economic status and living environment are recommended.

Key words: Log-linear model, Systematic sampling, hospital based surveillance

Introduction

Unhygienic environment is one of the major factors for spreading infectious diseases in the developing countries. Poor socioeconomic and the environment are profusely interrelated to infectious diseases. In many studies it has been shown that epidemics of diarrhea were due to poor hygiene and water sanitation. Diarrhea causes high mortality and morbidity of children less than five years mainly due to water pollution and deterioration of environmental

sanitation (Zeitlin et al., 1995). In rural Bangladesh, incidence of diarrhea was found to be highest among children of 10-12 months (Alam, 1995). This age group has greater exposure to environmental contamination and the other risk factors including fecal contaminations, garbage disposal in the open field and inadequate cleaning after defecation (Alam N et al., 1989).

Many intervention studies showed that the reduction of diarrhea up to 40% could be achieved through health education, domestic hygiene, maternal education, occupation and household size (Aziz, 1990). Studies conducted at community level showed that socio-economic factors are profoundly associated with slum dwelling (85%), less economic feasibility (41%) to buy soap for hand washing (Hoque, 2003). In West Bengal, a rural community study indicated that children could be protected significantly from diarrhea if mother's wrong concepts about diarrhea can be altered through educational intervention (Sircar et al., 1987). In Nigeria, street vendor's food played significant roll to flare the high risk of prolonged diarrhea among children (Ekanem et al., 1994). Socioeconomic status and exposure to health programs showed significant relationship with selected maternal preventive behaviors in the northeast Thailand (Thongkrajai et al., 1990).

Krishnan Rajendran is scientist of National Institute of Cholera and Enteric Diseases. His research interest is novel statistical models in cholera research. E-mail: rajenk20@yahoo.com. Thandavarayan Ramamurthy, Seiner Scientist of National Institute of Cholera and Enteric Diseases. E-mail: ramalmurthy@yahoo.com. Sujit Kumar Bhattacharya, renowned scientist in India, WHO, SEARO, New Delhi, India. E-mail: sujitkbhattacharya@yahoo.com.

A cross-sectional survey found association between mother's hygiene, socio economic status, household environment and mother's education level may contribute to the attitude of hand washing in Philippines (Sakisaka et al., 2002). From many investigations it is evident that a prime factor for environment is water born diseases all over the world. The Log –linear model should consider many categorical variable and environmental factors to generate useful information (Lakhan et al., 2002).

The categorical data can be highlighted the interrelationship with log linear analysis. The main emphasis of log linear analysis is that linear in the logarithms of expected frequencies of a contingency table that adequately describes or fits the associations and interaction which exist in the original frequency table. Log linear techniques are especially useful for frequency table with more than three variables when the number of possible associations and interactions among the variables became very large (Gibert et al., 1981). The main objective of this study is to analyze the socioeconomic and environmental factors in relation with diarrhea among children less than five years by log-linear model.

Materials and Method

Systematic sampling of every 5th hospitalized diarrhea case was made in two randomly selected days (24 hours) in a week from an active surveillance system at the Infectious Disease Hospital (IDH), Kolkata, India. Data were based on patient's information such as socio-economic status, demography, water source, sanitation and living environment, clinical outcome, patient management and laboratory diagnostic report. The IDH treats about 30,000 patients yearly as in and out patients from Kolkata city as well as semi-urban areas around the city.

Data were collected by pre-trained professional and manually edited before sending to data management center. The data were entered in the personal computer by two data entry professional in the pre-designed format profarma with inbuilt validation that run on Epiinfo (version 6.02) package. Data were randomly checked and matched for validation

and was finally analyzed by employing SPSS 4.0 ver. and S-plus software's.

Log Linear Model

Log linear model can be classified as non-standard or conventional. Non-standard models require specification of a set of hypotheses concerning the structure of the data. Non-standard log linear models are not hierarchical and therefore do not consider main effects and interaction used in conventional log linear models (Kennedy et al., 1983; Magidson et al., 1981; Breen, 1984; Rindskopf, 1990;). Conventional log linear models can be grouped into saturated and unsaturated models. A saturated log linear model fits the original frequency tables exactly. Unsaturated log linear models are also referred to as hierarchical log linear models (Bishop et al., 1975; Hagenaaers, 1990; Agresti, 1990). The model does not distinguish between independent and dependent variables. All variables are treated alike as response variables whose mutual association is to be explored. The model to represent the association between the variables

$$\ln F_{ij} = u + u_1(i) + u_2(j) + u_{12}(ij)$$

where the parameter $u_{12}(ij)$ represents the association between category i of variable 1 and category j of variable 2. $u_1(i)$ is the “main effect “ if category i^{th} of the row variable and $u_2(j)$ is the “main effect “ if category j^{th} of the column variable. This is known as a saturated model.

Environmental and socio-economic parameters were included to co-relate with diarrhea of less than five years age children. The independent variables were classified into three groups: socio-economic, resident pattern and sanitation and water source (Table 1). It can be assumed that the number of cases per cell has a multinomial distribution. The relationship between independent variable and occurrence of diarrhea in areas was studied employing Log linear Model. A backward procedure for log-linear modeling (SPSS- ver.4.0 software) was

LOG-LINEAR MODEL USING HOSPITAL BASED SURVEILLANCE

Table 1. Three Groups of Independent Variables

Socio-economic	Resident pattern & Sanitation	Water source
i) Family Income (Faminc)	i) Floor structure (Floor)	i) Source of drinking water (Drinkwat)
ii) No of children(<=5Yrs age) in the family(Child)	ii) Wall structure (Wall)	ii) Drink water from where (whrrhpvt)
iii) No of members in the family (Nomemb)	iii) No of rooms in the resident (Noroom)	iii) Type of storage (Typestor)
iv) Fathers occupation (Ocufoth)	iv) Place of defecation (Plcedefe)	iv) Water used for washing utensils (wasuten)
v) Mothers occupation (Ocumoth)	v) Place of disposal (Plcedisp)	v) Water used for bathing (Wtrbath)
vi) Education of father (Educf)		
vii) Education of mother (Educum)		

used to identify the best model. The independent variables significantly associated with the outcome variable from three groups were selected and pooled. The second order hierarchy model was constructed and analyzed. In the explained model, socio-economic, resident pattern & sanitation and water source were considered (Table 3).

Out of 4,601 diarrhoeal patients enrolled in this study, 1,119 cases were of children below 5 years of age. Using the distribution of the day frequency of diarrheal cases, the peak season of diffusion of diarrheal diseases was determined. The data was edited and scrutinized as two days in a week and determined mean distance of all patients those who were admitted at IDH in selected days of the week to know the distance from where the patients were coming that was fitted in the polar graph using the S-Plus software. The day frequencies are fitted into time series curve, which will explore exact existence of the distribution.

Final Models:

1) Socio-economic

Areas+Faminc+Ocufoth+Educf+Educum+Child+(Areas*Faminc)+(Areas*Ocufoth)

+(Areas*Educf)+(Areas*Educum)+(Areas*Child).

2) Resident Pattern & Sanitation

Areas+Noroom+Floor+Wall+Plcedisp+Plcedefe+(Areas*Noroom)+(Areas*Floor)+(Areas*Wall)+(Areas*Plcedisp)+(Areas*Plcedefe).

3) Water source

Areas+Whthpvt+Typestor+Wtrbath+(Areas*Whthpvt)+(Areas*Typestor)+(Areas*Wtrbath).

Results

Among the 1,119 children, 871(77.8%) and 248(22.2%) were from urban and rural areas respectively with the ratio of 3.5:1. Environment and economic status of patients were given importance. The distribution of socio-economic, resident pattern & sanitation and water source are presented in Table 2. Many variables were modified according to generated measure of central tendency of that variable. For example, occupation of the father was classified in to daily labors (e.g., day laborer, share cropper),

Table 2. Distribution of grouped variables with area wise diarrhea cases

S.No	Factors	Urban (n=871)		Rural (n=248)	
I	Socio-economic variables	No	%	No	%
1	Family Income				
	≤1000	189	21.7	108	43.5
	1001 - ≤2000	320	36.7	79	31.9
	2001 - ≤3000	268	30.8	46	18.5
2	>3000	94	10.8	15	6.0
	No of children (≤5yrs age) in the family				
	Single	545	62.6	137	55.2
	Double	214	24.6	56	22.6
	3-4 child	43	4.9	11	4.5
3	>4 child	69	7.9	44	17.7
	No of members in the family				
	1-3 members	232	26.6	63	25.4
4	4-6 members	525	60.3	154	62.1
	>6 members	114	13.1	31	12.5
	Fathers occupation				
	Daily labor	307	35.3	118	47.6
	Skilled workers	327	37.5	69	27.8
5	Owns petty business	158	18.1	43	17.3
	Others	79	9.1	18	7.3
	Mother occupation				
6	House wife	731	83.9	162	65.3
	Skilled workers	137	15.8	85	34.3
	Others	3	.3	1	.4
7	Father education status				
	School completed	512	58.8	131	52.8
	College & University	88	10.1	13	5.2
	Illiterate	271	31.1	104	42.0
8	Mother education status				
	School completed	455	52.2	102	41.1
	College & University	31	3.6	6	2.4
	Illiterate	385	44.2	140	56.5
9	Resident Pattern				
	Floor Structure				
	Non cemented	659	75.7	103	41.5
	Cemented	131	15.0	58	23.4
		66	7.6	69	27.8
		15	1.7	18	7.3

LOG-LINEAR MODEL USING HOSPITAL BASED SURVEILLANCE

10	Place of defecation				
	Sanitary	696	80.0	116	46.8
	Service	62	7.1	19	7.7
	Dughole	57	6.5	34	13.7
	Open field	56	6.4	79	31.8
11	No of Room in the house				
	≤2 Rooms	787	90.4	220	88.7
	3-4 Rooms	75	8.6	24	9.7
	>4 Rooms	9	1.0	4	1.6
12	Place of disposal				
	In to latrine	267	30.7	41	16.5
	In to drain	340	39.0	54	21.8
	In field	130	14.9	112	45.2
	Others	134	15.4	41	16.5
III	Water Sources				
13	Source of drinking Water				
	Tap	602	69.1	56	22.6
	Tube well	262	30.1	180	72.6
	Open well	7	.8	12	4.8
14	Drinking water from where				
	Private	233	26.8	54	21.8
	Common	635	72.9	185	74.6
	Others	3	.3	9	3.6
15	Type of storage				
	Earthen Pot (nm)	390	44.8	91	36.7
	Earthen Pot (wm)	197	22.6	98	39.5
	Bucket	257	29.5	56	22.6
	Others	27	3.1	3	1.2
16	Water used for washing utensils				
	Tap	401	46.0	32	12.9
	Tube well	149	17.1	94	37.9
	Pond	7	.8	6	2.4
	Open well	12	1.4	34	13.7
	Others	302	34.7	82	33.1
17	Water used for bathing				
	Tap	377	43.3	31	12.5
	Tube well	145	16.6	78	31.5
	Pond	7	.8	7	2.8
	Open well	24	2.8	41	16.5
	Others	318	36.5	91	36.7

skilled workers (e.g., boatman), small business owner (e.g., street vender, fisherman). Similarly other variables were modified (Table 2). The area wise results of the significantly associated independent variables from log-linear model of different groups are presented in Table 3.

In rural areas, family income of $\leq$ Rs 1000 ($p < 0.001$), father's occupation as daily labors ($p < 0.001$), illiterate mothers ($p < 0.003$) and father's education as an illiterate ($p < 0.001$) in Socio-economic group; non-cemented floor structure ($p < 0.001$), wall structure with mud ($p < 0.001$) and fecal disposal in field ($p < 0.001$) in resident pattern & sanitation group, and in water source category, type of storage with wide mouthed earthen pot ($p < 0.001$) and water used for bathing from open well and others ($p < 0.001$) were significantly associated with diarrhea. In urban areas, family income > 1000 but ≤ 2000 ($p < 0.001$), in socio-economic group; non-cemented floor structure ($p < 0.0001$), wall structure with brick ($p < 0.001$) place of defecation with sanitary ($p < 0.001$) and fecal disposal in latrine ($p < 0.001$), were significantly associated with diarrhea. In resident pattern & sanitation group no one was found significantly associated with diarrhea.

Discussion

Developing countries can reduce disease burden by 25% that is the equivalent to averting more than 9 million infant deaths by control and treatment of infectious diseases including malnutrition. The burden of diarrheal diseases was 92.8 million of Disability adjusted life years for fewer than 4 years children. Improving household environments could avert the annual loss of more than 80 million disability-free years of human life (World Bank, 1993). Environment and economic status are core factors for stabilizing the human life in health prosperity, particularly poor and unhygienic environment help in the spread of many infectious diseases in human. Different environmental factors have been proposed for the subsistence and spread of diseases. Previous community based studies conducted in Calcutta, showed that maternal behavior, feeding practices, living in non-cemented house, less family income, illiterate mother, wide mouth container for drinking water

storage, sharing common latrine, leftover food for next feeding, dirty baby cloths, less birth space between two children were responsible for childhood diarrhea (Ghosh et al., 1998).

The interest was focused on the use of log linear models to explore the data because this model is categorical and specially deals with environmental features. The log-linear model with broad spectrum of environmental factors previously has been used (e.g., Valtonen et al., 1994; Gonzalez et al., 1995; Badia et al., 1996). Log linear modeling permits to assess different factors such as for the socio-economic, resident pattern & sanitation and water source, which might influence diarrheal disease.

The independent variables have been grouped in to three viz. socio-economic, resident pattern and sanitation and water source, with at least 5 independent variables in each group viz. family income (Faminc), Floor structure (Floor), Source of drinking water (Drinkwat), etc. Faminc is formed a main effect and interaction association between two variables in the frequency table such as (Areas) and (Faminc), is denoted as Areas Faminc in the log linear model. By the definition of hierarchical log linear model, the term (Areas * Faminc) contains the two lower order main effect; similarly interaction among three variables can be analyzed in the log linear model.

The order interaction among three variables did not show any significant relation. Generally, various log linear models can be fitted to the data, the final log linear model that is selected need not contain all possible interaction in the frequency table since the purpose of log linear analysis is to find the simplest model that adequately describes the data. As a rule the greater the number of variables in the frequency table, the larger the number of possible log linear models that can be constructed.

Partial and marginal association tests and backward elimination procedure are normally used for selecting log linear models to be tested. The procedures of marginal and partial association are described to test main effects of three groups the log linear model is used (Dillon et al., 1984). All main effects are included in the base model, which is fitted to the original frequency table, and its level of significance is

LOG-LINEAR MODEL USING HOSPITAL BASED SURVEILLANCE

noted. Models are then fitted that omit each main effect in succession. The significance of the main effect variable is then determined by finding the difference between the chi-square values and the degrees of freedom of the base model. In a similar manner, all two-way, three-way interaction terms, and omitting each of the respective interaction term were conducted.

The log-linear model explored various factors associated with childhood diarrhea both in urban and rural community. In rural area, we found the socioeconomic factors like income of the family, father's occupation; parents' education and environmental factors like resident pattern, place of disposal of their children feces, water usage like drinking, bathing were significantly associated with diarrhea. In developing countries providing good living and socioeconomic conditions can prevent diarrheal diseases. In addition, many effective preventive measures should be adopted to reduce the diarrheal episodes.

In Australia, randomized controlled trial diarrhea on infection-controlled intervention has shown that 66% diarrheal episodes at the child care centers (Borghi et al., 2002; Roberts et al., 2000; Tompson, 1994). Compared to control group, hand-washing practice reduced 89% of diarrheal episodes in Indonesia (Wilson et al., 1991). In northern Pakistan, a case-control study revealed after village level implementation of water supply, sanitation awareness and practices about hygiene behavior as confounders were controlled 33% of diarrheal episodes at the early stage (Jensen et al., 2003; Nanan et al., 2003). In this study, July – October (monsoon to north monsoon) was peak season for arrival of maximum diarrheal cases. Studies conducted in China have also shown a similar trend in the occurrence of diarrhea (Yang et al., 1990). As shown in other studies, this study supports improvement of environment and economic status of the people is the first step in the prevention of diarrheal diseases.

Figure 1 (see Appendix) focuses the promotion of community education directly influence quality of human life. In this study, we have shown that in rural community less than five years old children were susceptible for diarrheal diseases owing to deterioration of personal hygiene of the parents and child which

are directly related with economic status of the family, education status of parents and water quality around them. It is known that in many developing countries the population is in progress nature due to which, eradication of infectious diseases particularly diarrhea is a hard task as lack of both education and economic status succumb the personal hygiene of every family. As shown in the hypothetical model (Fig-1), educate people, evaluate the economic status, emancipate poor environment might help to eradicate the infectious diseases in the rural community.

Acknowledgement

The authors thank Research fellows and Scientists for helping to generate the epidemiological data and staff of the Division of Clinical medicine, division of Epidemiology, National Institute of Cholera and Enteric Diseases and Infectious Diseases Hospital, Kolkata for enthusiastic participation in the active surveillance system.

Reference

- Agresti A. (1990). *Categorical data analysis*. New York: Wiley.
- Alam, N. (1995). Predictors of diarrhea in young Bangladeshi children. *Journal of tropical pediatrics*, 41(5), 278-280.
- Alam, N., Wojtyniak, B., Henry, F. J., & Rahaman, M. M. (1989). Mothers' personal and domestic hygiene and diarrhea incidence in young children in rural Bangladesh. *International Journal of Epidemiology*, 18(1), 242-247.
- Aziz, K. M., Hoque, B. A., Hasan, K. Z., Patwary, M. Y., Huttly, S.R., Rahaman, M. M., & Feachem, R. G. (1990). Reduction in diarrheal diseases in children in rural Bangladesh by environmental and behavioral modifications. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 84(3), 433-438.
- Badia J & Prouzet P. (1996). Loglinear models for studying associations between catches per unit effort and hydroclimatic variables. *Aquatic living Resources*, 9(3), 217-223.

- Bishop, Y. M. M., Fienberg, S. E., & Holland, P. W. (1975). *Discrete multivariate analysis: Theory and practice*. Massachusetts, Cambridge: MIT Press.
- Borghi, J., Guinness, L., Ouedraogo, J., & Curtis, V. (2002). Is hygiene promotion cost-effective? A case study in Burkina Faso. *Tropical Medicine & International Health*, 7(11), 960-969.
- Breen, F. H. (1984). Fitting nonhierarchical and association log linear models using GLIM. *Sociological Methods and Research*, 13, 77-107.
- Dillon, W. R., & Goldstein, M. (1984). *Multivariate analysis: Method and application*, NY: Wiley.
- Ekanem, E. E., Adediji, O. T., & Akitoye, C. O. (1994). Environmental and behavioral risk factors for prolonged diarrhea in Nigerian children. *Journal of Diarrheal Diseases Research*, 12(1), 19-24.
- Ghosh, S., Sengupta, P. G., Gupta, D. N., Mondal, S. K., Goswami, M., Bhattacharya, S. K., & Sircar, B. K. (1998). Maternal knowledge on risk behavioral practices and its association with diarrhea in a rural community of West Bengal, India. *The Journal of Communicable Diseases*, 30(4), 251-255.
- Gilbert, G. N. (1981). *Modeling society. An introduction to log linear analysis for social researchers*. London: George Allen and Unwin.
- Gonzalez, J. M. (1995). Modeling enteric bacteria survival in aquatic systems. *Hydrobiologia*, 316(2), 109-116.
- Hagenaars, J. A. (1990). *Categorical longitudinal data: Log linear panel, trend and cohort analysis*. Newbury Park, CA: Sage.
- Hoque, B. A. (2003). Hand washing practices and challenges in Bangladesh. *International journal of environmental health research*, 1, S81-S87.
- Jensen, P. K., Ensink, J. H., Jayasinghe, G., van der Hoek, W., Cairncross, S., & Dalsgaard, A. (2003). Effect of chlorination of drinking-water on water quality and childhood diarrhea in a village in Pakistan. *Journal of Health, Population, and Nutrition*, 21(1), 26-31.
- Kennedy, J. J. (1983). *Analyzing qualitative data: An introductory log-linear analysis of behavioral research*. NY: Praeger.
- Lakhan, V. C., & Lavalley, P. D. (2002). Use of loglinear models to assess factors influencing concern for the natural environment. *Environmental Management*, 30(1), 77-87.
- Magidson, J. J., Swan, H., & Berk, R. A. (1981). Estimating nonhierarchical and nested loglinear models. *Sociological Methods and Research*, 10, 3-49.
- Nanan, D., White, F., Azam, I., Afsar, H., & Hozhabri, S. (2003). Evaluation of a water, sanitation, and hygiene education intervention on diarrhea in northern Pakistan. *Bulletin World Health Organization*, May 16, 81(3), 160-165.
- Rindskopf, D. (1990). Nonstandard log linear models. *Psychological Bulletin*, 108(1), 150-162.
- Roberts, L., Jorm, L., Patel, M., Smith, W., Douglas, R. M., & McGilchrist, C. (2000). Effect of infection control measures on the frequency of diarrheal episodes in child care: a randomized, controlled trial. *Pediatrics*, 105(4 Pt 1), 743-746.
- Sakisaka, K., Wakai, S., & Wongkhomthong, S. A. (2002). Domestic hygiene behaviour of mothers with children aged 0-5 years old in Tayabo village, Nueva Ecija, the Philippines. *Asia-Pacific Journal of Public Health*, 14(2), 91-98.
- Sircar, B. K., Sengupta, P. G., Mondal, S. K., Gupta, D. N., Saha, N. C., Ghosh, S., Deb, B. C., & Pal, S. C. (1987). Effect of handwashing on the incidence of diarrhea in a Calcutta slum. *Journal of Diarrheal Diseases Research*, 5(2), 112-4.
- Thongkrajai, E., Thongkrajai, P., Stoeckel, J., Nanakhon, S., Karenjanabutr, B. & Sirivatanamethanont, J. (1990). Socioeconomic and health programme effects upon the behavioral management of diarrheal disease in northeast Thailand. *Asia-Pacific Journal of Public Health*, 4(1), 45-52.
- Thompson, S. C. (1994). Infectious diarrhea in children: controlling transmission in the child care setting. *Journal of Pediatrics and Child Health*, 30(3), 210-209.

LOG-LINEAR MODEL USING HOSPITAL BASED SURVEILLANCE

Valtonen, E. T., Tuuha, H., & Pugachev, O. N. (1994). Seasonal studies of the biology of the achtheres percarum in Perch, Perca Fluviatilis, from four Finnish lakes over a 3-year period. *Journal of Fish Biology*, 43(4), 621-632.

Wilson, J. M., Chandler, G. N., & Muslihatun, J. (1991). Hand-washing reduces diarrhoea episodes: a study in Lombok, Indonesia. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 85(6), 819-821.

World Bank. *World development report 1993- investing in health*. (1993). NY: Oxford University Press.

Yang, C. R., Meng, Z. D., Wang, X., Li, Y. L., Zhang, Y. X., & Zhao, Q. P. (1990). Diarrhoea surveillance in children aged under 5 years in a rural area of Hebei Province, China. *Journal of Diarrhoeal Diseases Research*, 8(4), 155-159.

Zeitlin, M. F., Ahmed, N. U., Beiser, A. S., Zeitlin, J. A., Super, C. M., & Guldán, G. S. (1995). Developmental, behavioral, and environmental risk factors for diarrhea among rural Bangladeshi children of less than two years. *Journal of Diarrheal Diseases Research*, 13(2), 99-105.

APPENDIX

Figure 1. Hypothetical model showing several intervention strategies in controlling diarrheal diseases. PS-Poor sanitation, PRP-Poor resident pattern, PQW-Poor quality water, PO-Parents Occupation, FI-Family income.

Brief Reports On the Length of NHL Shootouts

W. J. Hurley
Royal Military College of Canada

When NHL teams are tied after 60 minutes of regulation time and 5 minutes of sudden-death overtime, they go to a shootout to determine who gets the overtime point. Teams alternate shots until a winner is determined. The probability of observing shootouts of various lengths is calculated.

Key Words: probability tree, independent trials, Hockey shootout

Introduction

It is certainly true that statisticians have a fascination with records, particularly with sports records. The interested reader is referred to Nevzorov (2001) for a comprehensive treatment of the statistical theory associated with records. The primary motivation in this literature is the prediction of the way records will change over time. Noubary (2005) is illustrative of this genre.

When NHL teams are tied after 60 minutes of regulation time and 5 minutes of sudden-death overtime, they go to a shootout to determine who gets the overtime point. Teams alternate shots until a winner is determined. The shootout went 26 shots in a game between the Flyers and Rangers during the 2006-2007 season. The NHL record is 30 shots. Given the sudden death structure of shootouts, it seemed these ought to be rare events.

NHL Shootout Rules

The shootout is governed by Rule 89 (Note 4) of the NHL Rulebook: "Each team will be given three shots, unless the outcome is determined earlier in the shootout. After each team has taken three shots, if the score remains tied, the shootout will proceed to a "sudden

death" format. No player may shoot twice until everyone who is eligible has shot."

Rule 89 can be found at: <http://www.nhl.com/rules/rule89.html>. As Rule 89 suggests, it is possible to complete the shootout in fewer than six shots. For instance, if the team shooting first scores on its first two shots and the other team misses on its first two, there is no need for the teams to take their third shots. Once the shootout gets to the "sudden death" portion, the teams effectively generate a sequence of pairs of shots. The first pair comprises shots 7 and 8, the second, shots 9 and 10, and so on. Note that the shootout will proceed to pair $n+1$ shots if either both teams score on their pair n shots or both teams do not score on their pair n shots.

Shootout Statistics

Table 1 shows the frequencies of shootout lengths for the complete 2005-2006 season and the 2006-2007 season up to mid-November, 2006. In total, 176 games went to a shootout. Of these, 56 went to the sudden death format. Of some importance to the calculation are the probabilities that teams score on a shot. Over the short history of the shootout to date, 1,241 shots have been taken, and of these, 408 were successful. The frequency of scoring a goal, then, is

$$\frac{408}{1241} = 0.329$$

William Hurley is a Professor in the Department of Business Administration at the Royal Military College of Canada. His research interests are military operations research, decision analysis, and game theory. E-mail: hurley-w@rmc.ca.

Table 1.

Shootout Length (#shots)	Frequency (#games)
4	24
5	47
6	49
8	28
10	9
12	8
14	3
16	3
18	1
20	1
22	1
24	0
26	1
28	0
30	1

Table 2.

	<u>#Shots</u>	<u>#Goals</u>	<u>Frequency</u>
First 6 shooters	961	324	0.337
All other shooters	280	84	0.300
Totals	1,241	408	0.329

It is interesting to note that this frequency is comparable to the chance of hitting a Major League pitcher, a task generally considered to be one of the toughest in all of sport.

An obvious assumption is that all shots are independent trials and that each shooter has a probability, p , of scoring. However this ignores the fact that coaches will order shooters according to their ability to score. Hence it is probably the case that players shooting earlier are more likely to score. To assess these relative probabilities of scoring, I arbitrarily partitioned shooters into those taking the first 6 shots and

those taking the rest of the shots. Table 2 shows their relative performance over 1,241 shots. Note that the first 6 shooters have a higher frequency of scoring. However the difference is not enough to reject the null that both groups have the same probability of scoring.

Analysis

The first step is to calculate the probabilities of various length shootouts for the first 6 shots. Let p be the probability that a shooter scores. I assume this probability applies to each of the first three shooters on both teams. Let L_n be the probability that the length of the shootout is exactly n shots. Evaluating a probability tree, the following formulae are obtained. The probability of the shootout going to sudden death is

ON THE LENGTH OF NHL SHOOTOUTS

Table 2. Probability of Shootout Going to Sudden Death

<u>Length</u>	<u>Uniform</u>	<u>Non-uniform</u>
4	0.0974	0.1000
5	0.2466	0.2472
6	0.3185	0.3180
8	0.1490	0.1406
10	0.0832	0.0815
12	0.0465	0.0473
14	0.0260	0.0274
16	0.0145	0.0159
18	0.0081	0.0092
20	0.0045	0.0054
22	0.0025	0.0031
24	0.0014	0.0018
26	0.0008	0.0010
28	0.0004	0.0006
30	0.0002	0.0004

$$L_4 = 2p^2(1-p)^2$$

$$L_5 = 2p(1-p)[p^2 + (1-p)^2]$$

$$L_6 = 4p(1-p)\{[p^2 + (1-p)^2]^2 + p^2(1-p)^2\}$$

$$S_p = [p^2 + (1-p)^2]\{[p^2 + (1-p)^2]^2 + 6p^2(1-p)^2\}$$

Once in sudden death, the length of the shootout follows a geometric distribution. Note that $p^2 + (1-p)^2$ is the probability that the shooters either both score or both miss and $2p(1-p)$ is the probability that one scores and the other misses. Hence, the probability the shootout goes m shots for values of m at least 8 is

$$L_m = S_p [p^2 + (1-p)^2]^{m/2-4} 2p(1-p)$$

$$\text{for } m = 8, 10, 12, \dots$$

In the case where the first six shooters have a probability p of scoring and those shooting in the sudden death portion have a probability $\bar{p}$ of scoring, L_m is modified to

$$L_m = 2S_{\bar{p}} \bar{p}(1-\bar{p}) \left[\bar{p}^2 + (1-\bar{p})^2 \right]^{m/2-4}$$

$$\text{for } m = 8, 10, 12, \dots$$

The probabilities of shootout length for two assumptions about the underlying probabilities of scoring on a single shot are calculated by:

1. Uniform: All shooters score with probability $p = 0.329$.
2. Non-uniform: The first six shooters score with probability $p = 0.337$ and all other shooters score with probability $\bar{p} = 0.300$.

Table 3 gives the densities for lengths up to 30 shots. Not surprisingly the two distributions are close and the tail of the Non-uniform density is slightly larger.

Using the Non-uniform distribution, the probability of a shootout going at least 20 shots is 0.013. Of the 176 shootout games, 4 have gone at least 20 shots and the p -value of this outcome (the chance of observing 4 or more games taking at least 20 shots) is 0.19. Hence what actually happened is consistent with the theoretical prediction.

Another interesting question is the time between rare events. For instance, the average time between shootouts of at least 40 shots can be calculated. If the experience to date with penalty shots continues then the argument proceeds as follows. Using the Non-uniform probabilities, the chance of observing a shootout going at least 40 shots is 0.00005489. Assuming 1,230 NHL games a year and that, for each of these, there is a 12% chance that a game will get to a shootout, the average number of seasons between 40-shot shootouts is

$$\frac{1}{1230 \times .12 \times .00005489} = 123.4 \text{ seasons}$$

Equivalently, there is about a 50% chance of observing a shootout going at least 40 shots in the next 125 years.

References

Nevzorov, V. B. (2001). *Records: mathematical theory*. American Mathematical Society. Providence, RI.

Noubary, R. D. (2005). A procedure for prediction of sports records. *Journal of Quantitative Analysis in Sports*, 1(4). Available at <http://www.bepress.com/jqas/vol1/iss1/4>.

Statistical Software Applications & Review Robust General Linear Models and Graphics via a User Interface (Web RGLM)

Kimberly Crimin
Wyeth Research

Asheber Abebe
Auburn University

Joseph W. McKean
Western Michigan University

Rank-based procedures provide superior estimation and testing techniques when the data deviate from normality or contain gross outliers. However, these robust techniques are rarely incorporated in a nonparametric statistics or methods courses due to the lack of computational tools. One reason for this is the existence of certain unavoidable complexities in the numerical methods due to the absence of a closed-form solution for the rank estimation problem. This article introduces a user interface, Web RGLM, which may be used to perform rank-based analyses of linear models across the World Wide Web. These models include simple location problems to complicated ANOVA and ANCOVA designs with multiple comparison procedures. The robust and least squares analyses are presented side-by-side for immediate comparisons. Web RGLM meets many of the computational demands of the classroom as well as the computational demands of quantitative researchers. Several illustrative examples are provided.

Key words: R-estimation, RGLM, rank-based procedures, least squares, analysis of covariance

Introduction

Recent work on rank-based procedures for linear models has brought together a unified analysis of linear models analogous to the traditional analysis based on least squares. The rank-based analysis includes estimation, confidence procedures, testing of general linear hypotheses, and diagnostic methods. These rank-based analyses generalize the classical nonparametric rank procedures for one and two sample location

problems and they inherit the robustness and high efficiency of these simple methods. The recent article by McKean (2004) reviews this analysis while the monograph by Hettmansperger and McKean (1998) presents a thorough discussion of these rank-based analyses. Chapter 9 of the second edition of Hollander and Wolfe (1999) also offers a recent discussion of these methods. In Section 4, we give a quick overview of the rank-based analyses that are on our web page.

Kimberly Crimin is a Senior Principal Biostatistician in a Non-clinical Statistics group at Wyeth. She also works with robust procedures for discriminant analysis. Asheber Abebe is Associate Professor of Statistics at Auburn University. His research interests include robust regression and classification and exact simultaneous inference procedures. Joseph McKean is a Professor of Statistics. His research areas include nonparametric and robust statistics, along with computational algorithms for these procedures. He has co-authored three books in statistics. E-mail him at: joseph.mckean@wmich.edu

Traditional least squares analyses are based on estimation by minimizing the Euclidean (squared) norm, while the rank-based procedures are based on the minimization of a different norm. The minimization of this norm is a benign numerical problem which can be handled by existing numerical methods. However, to be of practical use these procedures must be easily computed. In this article we present an easy-to-use web version of these rank-based procedures. It allows the user to 'point-and-click' to perform these analyses for simple location problems through complex experimental designs. The output offers numerical results and diagnostic plots, produced by the R language; see Ihaka and Gentleman

(1996). Another advantage of the output is that it offers side-by-side comparisons of the robust and least squares (LS) analyses. If the analyses disagree then the user may choose to explore the data to determine reasons for this disagreement. This side-by-side comparison also serves as a very useful teaching tool. For instance, the student can immediately see the impact that perturbations of the data have on the LS and robust analyses.

Our web-based version of these analyses is discussed. Several examples are provided. It is found <http://fisher.stat.wmich.edu/slab/RGLM/>.

Web-Based RGLM

RGLM, (Robust General Linear Model), is the name of the FORTRAN program that performs the robust general linear model estimation and hypotheses testing described in Section 5. It was developed by Kapenga, McKean, and Vidmar (1988), and follows algorithms listed below. For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, the package RGLM returns a robust fit of this model. To make this program accessible to researchers, scientists and students, a web interface to RGLM was created. All the analyses discussed in this article were obtained using the Web based RGLM, which is available at <http://fisher.stat.wmich.edu/slab/RGLM/>.

The web interface to RGLM is module driven. Each module represents a different linear model that can be run using Web RGLM. Figure 1 is a screen capture of the home page for RGLM.

Note that many of the usual designs are given as options, from simple location models to complicated crossed factorial designs. When a user clicks on the name of the desired linear model (see Figure 1), a form is returned which allows the user to input the data and further customize the desired analysis. Some of the analysis options are: residuals and studentized residual plots, data plots appropriate for the model, contrasts along with type of interval, and type of scores used to estimate cell location (either Wilcoxon or signed-rank Wilcoxon). Each data input page describes the format of the input data set and contains an example data set. Data may be directly typed into the data entry

window or entered as a file that resides on the user's machine.

The freeware R (Ihaka & Gentleman, 1996) is used to produce the residual and data plots. Clicking the "Submit Data" button will result in a run of the desired analysis with the selected options. Clicking the "Clear Form" button will result in a default form and an empty data window.

For each module, both the traditional and rank-based analyses are provided. This summary has served as a useful teaching tool in applied nonparametric courses and methodology courses, in general. For a given data set, students can easily see if there is a difference in the analyses. In the case where the analyses differ, students can then try to determine why they differ by using residual plots and exploring the data to see if the discrepancy is caused by outliers or decidedly non-normal data, etc. It forces them to decide which analysis, if any, to use. Further, students can easily see how sensitive the robust and traditional analyses are by changing data points. For example, consider a one-sample problem. By repeatedly changing a data point, in a few seconds the student can have the data base to do comparison sensitivity plots of the Hodges-Lehmann estimator and the sample average.

The Web version of RGLM will run on any browser that is compatible with forms and, if the user selected residual or data plots, with graphics. All of the computations are done on the side of the server, reducing the hardware and software requirements of the user and ensuring uniformity of the output.

Behind the Scenes

The Web version of RGLM is a collection of CGI scripts, written in Perl (see Srinivasan (1997)), UNIX shell and FORTRAN programs. The statistical software R is used to obtain the user selected plots. The home page for Web RGLM and the input page for each linear model exist as separate HTML documents. The HTML page displaying the output is created by

Welcome to RANOVA - RGLM's ANOVA

Simply select the design that you have data for, input the data and click the submit button. The output will be generated for the hypotheses of interest. Each input page has an example data set that you can use as a test. The designs with covariates are towards the bottom of the page.

[One-way](#), performs a robust one-way analysis. Calculates the reduction in dispersion F-Statistic for the hypothesis that all the cell locations are the same.

[One Sample](#), calculates Signed Rank Wilcoxon test statistic and distribution free CI for θ the location parameter.

[Paired](#), calculates Signed Rank Wilcoxon test statistic and distribution free CI for θ the location parameter of the paired data.

[Simple Regression](#), performs robust regression. The Wilcoxon score function is used to determine the R-estimate of the slope and the intercept is estimated by the median of the residuals. Calculates the reduction in dispersion F statistic for the hypothesis the slope is to zero.

[Two-way](#), performs a robust two-way analysis. Calculates the reduction in dispersion F-Statistic for the test of interaction and main effects.

[Two Sample](#), calculates the Mann-Whitney-Wilcoxon test statistic and a distribution free CI for δ the difference in locations parameters.

[Additive](#), performs a robust two-way analysis assuming the factors do not interact. Calculates the reduction in dispersion F-statistic for the test of main effects.

[K-Way](#), performs a robust k-way analysis. Calculates the reduction in dispersion F-Statistic for the tests of interaction and main effects.

[Multiple Regression](#), performs robust multiple regression. The Wilcoxon score function is used to determine the R-estimate of the coefficients of the independent variables and the intercept is estimated by the median of the residuals. Calculates the reduction in dispersion F statistic for the hypothesis that all beta's are zero. In the future the user will be able to specify a subset of the betas.

[RGLM Format Page](#), allows for user input of design matrix and hypotheses matrices. These have to be in RGLM format. For a pdf version of a mini-manual [Click Here](#)

Analysis of Covariance Models

[Oneway](#), performs a robust oneway with covariates analysis. Calculates the F-Statistic for the hypotheses that all the cell locations are the same and that the covariates are zero.

[Twoway](#), performs robust two-way with covariates analysis. Calculates the reduction in dispersion F-Statistic for tests of main effects, interaction and covariate effect.

[Additive](#), performs robust two-way with covariates analysis assuming the factors do not interact. Calculates the reduction in dispersion F-Statistic for test of main effect and covariate effect.

[K-way](#), performs robust k-way analysis. Calculates the reduction in dispersion F-Statistic for tests of main effect, covariate effect and interactions.

Figure 1: RGLM Home Page

the CGI script once RGLM has executed. This section provides a brief overview of the behind the scenes workings of Web RGLM.

RGLM is the main FORTRAN program that performs the robust analysis. RGLM requires three input files of a specific format. One file contains options for the rank-based analysis, another file contains the $X|Y$ augmented matrix, where X is the design matrix and Y is the data matrix, and the third file contains the hypotheses matrices. To shorten the learning curve for the user and to make Web

RGLM 'point-and-click', a FORTRAN program creates the three RGLM input files from the information provided by the user in the data input page.

The use of a FORTRAN program to create the input files also provides some security for the server, since the form only sends data and options to the CGI program and not commands. Within the CGI program, the data is checked to make sure it only contains digits. If characters other than digits are found, then an error page is returned to the user indicating an

Figure 2: Covariate vs. Final Distance by Treatment. The solid line is the robust fit, while the dashed line is the LS fit. Plots are for: Control, Upper Left Panel; Treatment 1, Upper Right Panel; Treatment 2, Lower Left Panel; and Treatment 3, Lower Right Panel.

Figure 3: Residual Plots

error in the data input file. The FORTRAN program that creates the RGLM input files is the only program that is run with a system call that uses input provided by the user. All other system calls are executed on data files created by this FORTRAN program. This strategy limits the number of doors left open to the server.

The RGLM program allows the user to specify the name of the input file containing the augmented $X|Y$ matrix and the hypothesis matrix, but does not allow the user to specify the names of the output files. To allow multiple users to run Web RGLM at the same time without clobbering each other, each user is assigned a user ID. The user ID maps to a temporary directory and all files created for that run are stored in the temporary directory. After the HTML page containing the output is returned, all files in the temporary directory are removed along with the temporary directory. If a user runs multiple analyses in the same web session, a temporary directory is created and removed on each run. An earlier version of Web RGLM stored the user ID as a cookie. In this previous version, the temporary directories were removed after a prescribed length of time. This caused unwanted complications when a web session exceeded the allowable time.

When a user selects data plots or residuals plots, the CGI program writes the R code to create the plots to a file. Then R is run in batch mode, producing a postscript file containing the plots. To display the plots in an HTML page, the postscript file is converted to a gif file using Netpbm graphics utilities available at sourceforge.net/projects/netpbm.

Examples

Using the Web version of RGLM we offer three illustrative examples of the rank-based analysis, comparing it with the traditional Least Squares analysis in each case. We use the default Wilcoxon scores. These scores are based on a linear score function (see Section 5.2) and for the one and two-sample location problems these scores result in the usual Mann-Whitney-Wilcoxon analyses. They require no tuning constants. Other scores can be used, as briefly cited in Section 4.

Analysis of Covariance: Snake Data Set

The dataset used for this example was discussed in Afifi and Azen (1972). The purpose of the experiment was to compare methods of reducing human's fear of snakes. There are three methods intended to reduce ones fear of snakes and one control, or placebo. Forty subjects were randomly assigned to the four treatments. To measure ones fear of snakes, a behavior approach test was used to determine how close one could walk towards a snake without feeling uncomfortable. The behavior approach test was given to each subject before and after treatment. The score on the before treatment test was taken as a covariate.

To obtain the rank-based analysis of this data set using Web RGLM proceed as follows: from the home page, click on "Oneway" under Analysis of Covariance Models (see Figure 1) and drop the data and covariate into the data boxes. For this analysis, we included covariate by treatment interaction in the model and used cell medians as the estimates of location.

There are several options for plots available to the user. For the analysis below, we requested covariate versus response by treatment, residuals versus fitted values and a normal $q-q$ plot of the residuals. These plots are shown in Figure 2 and Figure 3, respectively. The residual plot indicates that the data are heteroscedastic which can be eliminated by the square root transformation applied to the response variable.

Figure 4 contains the analysis part of the output from Web RGLM. It is clear from the plots of the response, final distance, by treatment, Figure 2, that the treatment slope parameters are not the same. The comparison analyses show that the robust F -test for parallelism detects this difference with a p value of 0.01, but that the LS F -test with p value is 0.09 fails to detect this difference at the 5% level. Based on the $q-q$ plot of residuals, Figure 3, the underlying error structure appears to be heavy tailed, so the difference in the analyses is not surprising.

One-Way Analysis: Creatine Data Set

For our second example we have chosen a data set from a pharmaceutical study. The data

RGLM's ANOVA Output

One Way with Covariates Output

Estimates of Cell Location and Standard Errors

	Wilcoxon Rank Based $\tau\text{-hat}=3.94311$				Cell Means $\sigma\text{-hat}=5.81833$			
	Estimate	SE	t-ratio	p-value	Estimate	SE	t-ratio	p-value
FinalDistance-1	27.2819	3.62576	7.52447	1.44004e-08	25.6149	5.26431	4.86577	2.93435e-05
FinalDistance-2	-1.78198	2.77338	-0.64253	0.525106	-1.38736	3.97956	-0.34862	0.729658
FinalDistance-3	-6.67944	2.43761	-2.74015	0.00995914	-6.39363	3.46804	-1.84358	0.0745225
FinalDistance-4	2.85244	2.39913	1.18895	0.243208	7.80823	3.40912	2.2904	0.0287341
Cov*FinalDistance-1	-0.0242747	0.195477	-0.124182	0.901948	0.0693046	0.288439	0.240274	0.81165
Cov*FinalDistance-2	0.825997	0.149463	5.52643	4.29335e-06	0.83046	0.220543	3.76552	0.000673494
Cov*FinalDistance-3	0.867943	0.118588	7.319	2.5455e-08	0.868668	0.174984	4.96427	2.20457e-05
Cov*FinalDistance-4	0.660261	0.127124	5.19382	1.13076e-05	0.489658	0.187581	2.61039	0.0136462

Coefficients of
Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.797625	0.76295

Hypothesis: Cell Locations are the same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
FinalDistance effect	12.662	3,32	1.26958e-05	9.62872	3,32	0.000111222

Hypothesis: Slopes are the Same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
Cov*FinalDistance effect	4.06897	3,32	0.0147894	2.33664	3,32	0.0922737

Figure 4: Screen Capture of Rank-Based Analysis of Snake Data

set contains the results of an experiment that was run on mice to determine the effects of different doses of an experimental compound on the amount of creatine cleared from the body. The mice were randomly divided into six groups. The first group formed the control which had a dose level of 0 of the compound. The other five groups each had a different dose of the compound. The data have been corrected for the body weights of the mice. Thus the appropriate design is a one-way design. Besides the test of an overall effect, it was of interest to compare the five groups to the control. On the RGLM page, "Oneway" was selected.

One of the plots we checked on the form was the comparison boxplots of the levels which is shown in Figure 5. Besides the apparent outliers, this plot indicates that all the treatment levels may be significantly different from the placebo.

For the contrast query on the RGLM one-way page, we checked *versus control* and entered the level (1) for the control. We selected the Tukey-Kramer multiple comparison procedure, (MCP). As shown in Figure 6, the Wilcoxon ANOVA detects these differences, the *F*-statistic has the value 7.24 with *p*-value 0.00001. In contrast, note that the LS *F*-statistic has *p*-value 0.056. The outliers impaired its

Figure 5: Comparison Boxplots of the observations by level and q - q plot of Wilcoxon Studentized Residuals for Creatine Data

Hypothesis: Cell Locations are the same

	Wilcoxon R			Least Squares		
	F-Statistic	df	p-value	F-Statistic	df	p-value
trt effect	7.23668	5,80	1.23382e-05	2.26084	5,80	0.0561732

Tukey-Kramer

Contrast	Wilcoxon R Contrasts			Least Squares Contrasts		
	Estimate	Error Term	Interval	Estimate	Error Term	Interval
2 1	2.36633	1.75528	(0.61105,4.12161)	2.75157	5.16927	(-2.4177,7.92083)
3 1	2.5935	1.64005	(0.953452,4.23355)	5.17179	4.82991	(0.341872,10.0017)
4 1	2.42003	1.601	(0.819034,4.02103)	3.92346	4.7149	(-0.791448,8.63836)
5 1	2.37465	1.601	(0.773654,3.97565)	3.0671	4.7149	(-1.64781,7.782)
6 1	2.34548	1.61942	(0.72606,3.9649)	2.463	4.76916	(-2.30617,7.23216)

Figure 6: Wilcoxon and Least Squares ANOVAs for Creatine Data

power. The table in Figure 6 summarizes the MCP study. For the Wilcoxon analysis, the Tukey-Kramer procedure declares that all five levels differ significantly from the control while

the LS version of the Tukey-Kramer procedure only declares that the third level differs significantly from the control.

Multiple Regression: Snow Geese Data Set

In this example, we consider the snow geese data set discussed on page 441 of Hollander and Wolfe (1999). It is a multiple regression problem with four predictors. The response is the time, minutes before (-) or after (+) sunrise, that lesser snow geese leave their overnight roost sites to fly to their feeding areas. The predictors are: x_1 , the air temperature in Celsius; x_2 , relative humidity; x_3 , light intensity; and x_4 , percent cloud cover. Data were collected for $n=36$ days. We assume the linear model,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i, \\ i = 1, 2, \dots, 36. \quad (1)$$

Besides estimating the regression coefficients, the following two hypotheses are of interest:

$$H_{01} : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \quad (2)$$

$$H_{02} : \beta_1 = \beta_2 = 0 \quad (3)$$

Hollander and Wolfe used the *rregr* command of Minitab to perform this analysis. We show how it is easily performed by the RGLM web page.

On the web page (Figure 1), click on "Multiple Regression". Next, drop in the data in the form $\mathbf{X} \ Y$ into the data box. The user has a choice on the estimate of the intercept, either the median of the residuals or the Hodges-Lehmann estimate of location based on the residuals. The hypothesis H_{01} is the usual regression hypothesis that all regression coefficients are zero, except for the intercept. Web RGLM always obtains the robust and LS tests for this hypothesis. For the second hypothesis, H_{02} , the reduced model is $Y_i = \beta_0 + \beta_3 x_{i3} + \beta_4 x_{i4} + e_i$. To obtain the test of H_{02} , indicate that this reduced model is to be fit by entering 3 and 4 into the box labeled "Enter column ids, 1 to p , to include in the reduced model."

Figure 7 shows the output. The full model estimates are given in the first table. The robust and LS fits are similar, except for the

estimate of β_1 in which the fits differ by about a half of a standard error. This may have been caused by the one outlier in the data set as seen in the residual and $q-q$ plots of the robust fit as shown in Figure 8. The tests that all regression coefficients are 0, H_{01} , are given in the third table, while the tests of H_{02} are given in the last table. This later hypothesis concerns dropping β_1 . As with the estimate of β_1 , the robust F test is more significant than the LS F test.

Conclusion

The statistical computation tool introduced in this article uses state-of-the-art web interfacing to provide users access to robust nonparametric methods. In addition to the traditional ASCII text output provided by RGLM, Web RGLM provides graphics for visual assessment of the data and model diagnostics. Graphics associated with rank-based procedures have customarily been produced using other statistical software after the output from RGLM is manually edited. With the web interface available, this cumbersome activity is now unnecessary. Moreover, the user is not limited to specific score functions. The RGLM Format page gives the user the option of choosing a score function, in addition to several other options, thus, retaining the flexibility of RGLM. There is an online manual describing customized analyses which the user can download.

There is an experimental companion to Web RGLM that uses high breakdown (HBR) techniques. This can be found at the URL: <http://fisher.stat.wmich.edu/slab/RGLM/HBR2>. As with the Web RGLM page, it offers side-by-side comparisons of the high breakdown and LS fits. These techniques, developed by Chang et al. (1999), use a stochastically weighted Wilcoxon norm to obtain estimators that are robust to outliers in both design and response space, while the Wilcoxon analysis is only robust in response space. We plan on finishing this page in the future. Also, we are planning future expansions of the page to other designs, including nested designs, generalized estimating equations, nonlinear models, and mixed models.

ROBUST GENERAL LINEAR MODELS AND GRAPHICS VIA A USER INTERFACE

Multiple Regression Output

Parameter Estimates and Standard Errors

	Wilcoxon R tau-hat=8.42223				Least Squares sigma-hat=8.09161			
	Estimate	SE	t-ratio	p-values	Estimate	SE	t-ratio	p-values
Intercept	-51.4134	9.19765	-5.58984	3.95836e-06	-52.9939	8.78728	-6.03076	1.12429e-06
Beta 1	1.03877	0.275392	3.77196	0.000685556	0.912981	0.264581	3.45066	0.00163517
Beta 2	0.126197	0.118441	1.06549	0.294883	0.142532	0.113791	1.25257	0.219729
Beta 3	2.53563	0.781934	3.24277	0.00283114	2.516	0.751238	3.34914	0.00214101
Beta 4	0.089625	0.0457174	1.96041	0.058988	0.0922053	0.0439228	2.09926	0.0440324

Coefficients of Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.692526	0.758723

Hypothesis: Model does not fit

Wilcoxon R			Least Squares		
F-Statistic	df	p-value	F-Statistic	df	p-value
17.4554	4,31	1.19209e-07	24.3707	4,31	0

Reduced Model Parameter Estimates and Standard Errors

	Wilcoxon R tau-hat=9.82016				Least Squares sigma-hat=9.29604			
	Estimate	SE	t-ratio	p-values	Estimate	SE	t-ratio	p-values
Intercept	-54.931	6.89528	-7.96646	2.78399e-09	-55.1992	6.53087	-8.45205	7.18156e-10
Beta 3	4.11602	0.621827	6.61924	1.3675e-07	4.37963	0.588639	7.44026	1.24672e-08
Beta 4	0.177072	0.0446428	3.96641	0.000356513	0.176112	0.0422602	4.16734	0.000199932

Reduced Model Coefficients of Determination

Wilcoxon R	Least Squares
R2 (robust)	R-squared
0.595624	0.661004

Hypothesis: User Specified Reduced Model

Wilcoxon R			Least Squares		
F-Statistic	df	p-value	F-Statistic	df	p-value
6.57319973451212	2,31	0.00417089	6.27759176687857	2,31	0.0051403

Figure 7: RGLM's ANOVA Output for the Snow Geese Data

Figure 8: Wilcoxon studentized residual and q - q plots and LS ANOVAs for the Snow Geese Data

Background

Just like the traditional least squares procedures, rank-based procedures give a unified approach to testing and estimation problems. The recent monograph by Hettmansperger and McKean (1998) ('HM' hereinafter) gives a detailed treatise of rank-based procedures for handling problems of estimation and testing in situations ranging from the simple one sample location problems to the highly complicated multi-factor experimental designs. In this section we briefly review rank-based estimation and testing procedures and direct the reader to HM for further details.

Linear Models

Let $\mathbf{Y} = (Y_1, \dots, Y_n)'$ denote the $n \times 1$ vector of observations which follows the linear model

$$\mathbf{Y} = \mathbf{1}\alpha + \boldsymbol{\eta} + \mathbf{e}, \boldsymbol{\eta} \in \Omega \quad (4)$$

where $\mathbf{1}$ is an $n \times 1$ vector of ones, Ω is a subspace of $\mathbf{R}^n$ spanned by the columns of a centered $n \times p$ design matrix $\mathbf{X}$, and $\mathbf{e}$ is an $n \times 1$ vector of random errors.

In addition to estimating α and $\boldsymbol{\eta}$ we test general linear hypotheses such as

$$H_0 : \boldsymbol{\eta} \in \omega \text{ versus } H_A : \boldsymbol{\eta} \in \Omega \cap \omega^\perp \quad (5)$$

where $\omega \subset \Omega$ is $p - q$ dimensional for $0 \leq q \leq p$. In the following we shall refer to the model given in (4) as the full model and the same model under H_0 as the reduced model.

R-Estimation

The estimate of $\boldsymbol{\eta}$ will be obtained by minimizing the distance between $\mathbf{Y}$ and the space Ω . The distance we minimize for R -estimation is based on the R pseudonorm defined as

$$\|\mathbf{u}\|_\phi = \sum_{i=1}^n a(R(u_i))u_i, \mathbf{u} \in \mathbf{R}^n \quad (6)$$

where $R(u_i)$ denotes the rank of u_i among $u_1, \dots, u_n$, and $a(i) = \phi(i/(n+1))$ for some nondecreasing score function ϕ defined on the interval $(0,1)$ and standardized such that

$\int \varphi = 0$ and $\int \varphi^2 = 1$. For the proof that (6) is indeed a pseudonorm the reader is referred to McKean and Schrader (1980). The set $\{a(1), a(2), \dots, a(n)\}$ is called the set of rank scores. The most common R scores used in practice are the Wilcoxon scores which are generated by $\varphi(u) = \sqrt{12}(u - 0.5)$; i.e a linear score function. In the simple location models, the rank-based analyses based on this score function are the Mann-Whitney-Wilcoxon procedures. The L_1 pseudonorm is another popular special case of (6) obtained when $\varphi(u) = \text{sgn}(u - 0.5)$. In the location cases, analyses based on the sign scores are the median (Mood) procedures.

The R -estimator of $\boldsymbol{\eta}$ is a vector $\hat{\mathbf{Y}}_\varphi$ such that

$$\|\mathbf{Y} - \hat{\mathbf{Y}}_\varphi\|_\varphi = \min_{\boldsymbol{\eta} \in \Omega} \|\mathbf{Y} - \boldsymbol{\eta}\|_\varphi \equiv D(\Omega) \quad (7)$$

The R -estimates are analogous to the least squares estimates. Suppose we use the Euclidean norm $\|\mathbf{u}\|_{LS}^2 = \sum (u_i - \bar{u})^2$. There the estimator is, of course, $\hat{\mathbf{Y}}_{LS} = \mathbf{H}\mathbf{Y}$ where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix onto the column space of the centered design matrix $\mathbf{X}$. To obtain the R -estimates we simply replace the Euclidean norm by the norm given in (6).

Estimation of Regression Coefficients

Rewriting (4) as $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector, the R -estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}_\varphi$, is the solution vector of the p normal equations

$$\mathbf{X}\hat{\boldsymbol{\beta}}_\varphi = \hat{\mathbf{Y}}_\varphi \quad (8)$$

Based on $\hat{\boldsymbol{\beta}}_\varphi$, we estimate the intercept as

$$\hat{\alpha}_S = \text{med}\{Y_i - \mathbf{x}'_i \hat{\boldsymbol{\beta}}_\varphi\} \quad (9)$$

Assume that the random errors follow a distribution G with density g and median $\theta_e = G^{-1}(1/2)$. Under some mild regularity conditions

$$\begin{pmatrix} \hat{\alpha}_S \\ \hat{\boldsymbol{\beta}}_\varphi \end{pmatrix} \text{ has an approximate } N_{p+1} \left(\begin{bmatrix} \alpha \\ \boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} n^{-1}\tau_S^2 & \mathbf{0}' \\ \mathbf{0} & \tau_\varphi^2 (\mathbf{X}'\mathbf{X})^{-1} \end{bmatrix} \right) \quad (10)$$

distribution, where

$$\begin{aligned} \tau_\varphi &= \left[\int \varphi(u) \varphi_g(u) du \right]^{-1} \\ \tau_S &= [2g(\theta_e)]^{-1}, \text{ and} \\ \varphi_g(u) &= -[g(G^{-1}(u))]^{-1} [g'(G^{-1}(u))] \end{aligned} \quad (11)$$

Thus we have an asymptotic $100(1 - \gamma)\%$ confidence interval for the linear combination $\mathbf{l}'\boldsymbol{\beta}$ given by

$$\mathbf{l}'\hat{\boldsymbol{\beta}}_\varphi \pm t_{(\gamma/2, n-p-1)} \hat{\tau}_\varphi \sqrt{\mathbf{l}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{l}} \quad (12)$$

where $\hat{\tau}_\varphi$ is an estimate of τ_φ obtained as in Koul *et. al.* (1987), briefly discussed below.

Estimation of Scale

Let $\hat{\mathbf{e}}$ denote the vector of residuals based on the R -fits and let $\varsigma = \tau_\varphi^{-1}$. Then from (11) it follows upon integrating by parts that

$$\varsigma = \int_{-\infty}^{\infty} g(x) d\varphi(G(x)) \quad (13)$$

The estimate of $g(x)$ is obtained using the rectangular kernel density estimator

$$\hat{g}_n(x) = (2nt_n)^{-1} \sum_{i=1}^n I(|x - \hat{e}_i| \leq t_n)$$

where $2t_n$ is the window width which will be decided later and $I(A)$ is the indicator function of the event A . Hence an estimate of ζ is,

$$\hat{\zeta} = (2nt_n)^{-1} \left\{ \sum_{j=1}^n \sum_{i=1}^n (\varphi(j/(n+1)) - \varphi(j/(n+1))) I(|e_{(j)} - e_{(i)}| \leq t_n) \right\}$$

where $\hat{e}_{(j)}$ denotes the j th ordered residual. Using the mean value theorem, standardize the expression in braces above as

$$H_n(z) = (n^2 \{(\varphi(1) - \varphi(0))c\})^{-1} \sum_{j=1}^n \sum_{i=1}^n \varphi'(R(\hat{e}_j/(n+1))) I(|\hat{e}_j - \hat{e}_i| \leq z)$$

The constant c is chosen so that H_n is an empirical distribution function of the absolute differences $|\hat{e}_j - \hat{e}_i|$. Then choose $t_{n\delta}$ so that $H_n(t_{n\delta}) = 1 - \delta$ for $0 \leq \delta \leq 1$. Our estimate of ζ is then,

$$\hat{\zeta} = \frac{H_n(t_{n\delta}/\sqrt{n})(\varphi(1) - \varphi(0))}{2t_{n\delta}/\sqrt{n}} \quad (14)$$

Thus our estimate of τ_φ is given by

$$\hat{\tau}_\varphi = \hat{\zeta}^{-1}$$

Koul et al. (1987) showed that this estimate is consistent for τ_φ under both symmetric and asymmetric error distributions.

Testing

Testing the hypothesis given in (5) will be performed using an F -type test statistic given by

$$F_\varphi = \frac{[D(\omega) - D(\Omega)]/q}{\hat{\tau}_\varphi/2} \quad (15)$$

where $D(\omega) \equiv \min_{\boldsymbol{\eta} \in \omega} \|\mathbf{Y} - \boldsymbol{\eta}\|_\varphi$ is the minimum dispersion under the restriction imposed by H_0 . The quantity qF_φ has an

asymptotic χ^2 distribution. Small sample studies, however, indicate that F should be compared to F distribution critical values with q and $n - p$ degrees of freedom.

Algorithm

Consider the QR-decomposition of $\mathbf{X}$

$$\mathbf{Q}'\mathbf{X} = \mathbf{R} \quad (16)$$

where $\mathbf{R}$ is an $n \times p$ upper triangular matrix of rank p and $\mathbf{Q}$ is an $n \times n$ orthogonal matrix. We may write $\mathbf{Q}$ as $[\mathbf{Q}_1 \mathbf{Q}_2]$ where $\mathbf{Q}_1$ is an $n \times p$ matrix whose columns form an orthonormal basis for the column space of $\mathbf{X}$. We can now write the k th Newton step as

$$\hat{\mathbf{e}}^{(k)} = \hat{\mathbf{e}}^{(k-1)} - \hat{\tau}_\varphi \mathbf{H}\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)})) \quad (17)$$

where $\mathbf{H} = \mathbf{Q}_1 \mathbf{Q}_1'$ and $\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)}))$ is a vector whose j th component is $a(R(\hat{e}_j^{(k-1)}))$. Here is the formal algorithm. Let ε_D be a given tolerance.

Step 0: Set $k=1$. Obtain initial residuals $\hat{\mathbf{e}}^{(k-1)}$, $\hat{\tau}_\varphi^{(k-1)}$, and the $(k-1)$ th step dispersion, $D^{(k-1)}$.

Step 1: Get $\hat{\mathbf{e}}^{(k)}$ as in (17). Obtain $\hat{\tau}_\varphi^{(k)}$, and $D^{(k)}$.

- If $D^{(k)} < D^{(k-1)}$, then go to Step 2.
- Else perform a linear search (see HM pp. 186-187) along the direction $\hat{\tau}_\varphi \mathbf{H}\mathbf{a}(R(\hat{\mathbf{e}}^{(k-1)}))$ for a value which minimizes D , then go to Step 2.

Step 2: If $[D^{(k-1)} - D^{(k)}]/D^{(k-1)} < \varepsilon_D$, then go to Step 3. Otherwise set $\hat{\mathbf{e}}^{(k-1)} = \hat{\mathbf{e}}^{(k)}$ and go to Step 1.

Step 3: Obtain estimates as $\hat{\mathbf{Y}} = \mathbf{Y} - \hat{\mathbf{e}}^{(k)}$, $\hat{\tau}_\varphi = \hat{\tau}_\varphi^{(k)}$, and $\hat{\boldsymbol{\beta}}$ by solving $\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}}$.

As a final note we mention that the QR-decomposition can be used to form reduced model design matrices for testing the hypotheses in (5) (see Theorem 3.7.2 of HM).

References

- Abebe, A. , Crimin, K. , McKean, J. W., Haas, J. V. & Vidmar, T. J. (2001), Rank-based procedures for linear models : Applications to pharmaceutical science data. *Drug Information Journal*, 35, 347-371.
- Afifi, A. A. & Azen, S. P. (1972), *Statistical analysis: A computer oriented approach*. New York: Academic Press.
- Chang, W. H., McKean, J. W., Naranjo, J. D., & Sheather, S. J. (1999), High breakdown rank regression, *Journal of the American Statistical Association*, 94, 205-219.
- Hettmansperger, T. P., & McKean, J. W. (1998), *Robust nonparametric statistical methods*, London: Arnold.
- Hettmansperger, T. P., McKean, J. W., & Sheather, S. J. (2000), Robust nonparametric methods, *Journal of the American Statistical Association*, 95, 1308-1312.
- Hollander, M., & Wolfe, D. A. (1999), *Nonparametric statistical methods, Second edition*, New York: Wiley.
- Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5, 229-314.
- Kapenga, J. A., McKean, J. W., & Vidmar, T. J. (1995), *RGLM: Users manual*, version 2, *SCL Technical Report*, Dept. of Statistics, Western Michigan University.
- Koul, H. L. Sievers, G. L. & McKean, J. W. (1987), An estimator of the scale parameter for the rank analysis of linear models under general score functions, *Scandinavian Journal of Statistics*, 14, 131-141.
- McKean, J. W. (2004), Robust analyses of linear models, *Statistical Science*, 19, 562-570.
- McKean, J. W., & Schrader, R. (1980), The geometry of robust procedures in linear models, *Journal of the Royal Statistical Society, Series B*, 42, 366-371.
- McKean, J. W., Vidmar, T. J., & Sievers, G. L. (1989), A robust two stage multiple comparison procedure with application to random drug screen, *Biometrics*, 45, 1281-1297.
- Srinivasan, S. (1997), *Advanced PERL programming*, Sebastopol, CA: O'Reilly.

Effect on Recreation Benefit Estimates from Correcting for On-Site Sampling Biases and Heterogeneous Trip Overdispersion in Count Data Recreation Demand Models (STATA)

Roberto Martínez-Espíñeira
Memorial University of Newfoundland

Joseph M. Hilbe
Arizona State University
NASA/Jet Propulsion Laboratory
California Institute of Technology

Correction procedures (STATA commands NBSTRAT and GNBSTRAT) are applied to simultaneously account for zero-truncation, endogenous stratification, and overdispersion, and also consider heterogeneity in the overdispersion parameter. Their effect is shown on welfare estimates from previous studies, confirming that the routines perform the appropriate correction and only when endogenous stratification is expected.

Key words: on-site sampling, heterogeneous overdispersion, zero-truncation, endogenous stratification, count data

Introduction

When analysing and predicting individual demand and behaviour in a variety of settings, researchers often resort to count data models to handle the special characteristics of the dependent variable and they often collect the data on-site for reasons of cost-effectiveness. This is the case, for example, of many recent recreational demand studies based on the travel cost method (TCM). The TCM is used to value public areas used for recreational purposes that require most users to travel to the site (Braden & Kolstad, 1991; Freeman, 1993). The TCM

assumes that the costs individuals incur when travelling to the site can act as surrogate prices for their recreational experience and that the number of trips to the site should decrease with increases in distance travelled and other factors that increase the total travel cost. Exploiting this empirical relationship between increased travel cost and declining visitation rates makes it possible to estimate a demand relationship. This demand schedule can be used to estimate the total benefits derived by visitors (e. g. consumer surplus). A similar approach can be applied in a variety of settings related to individual demand and behaviour analysis, but we will focus here on empirical applications of the individual single-site travel cost method. In a sense, the single-site TCM could be seen as a gravity model restricted to one destination from which no departures originate.

Roberto Martínez-Espíñeira is an Associate Professor of Economics. His research focuses on the econometric analysis of travel cost and contingent valuation data used for environmental valuation. E-mail him at rmartinezesp@mun.ca. Joseph M. Hilbe is a Solar System Ambassador with NASA/Jet Propulsion Laboratory, an Adjunct Professor of Statistics at Arizona State University, and an Emeritus Professor at the University of Hawaii. He is an elected Fellow of the American Statistical Association and the International Statistical Institute.

Gravity models are popular among geographers and transportation analysts and have also been used by recreation planners/economists to distribute regional recreation use across sites. However, they are somewhat less popular with economists. Economists typically work with the visitation-origin data to predict visitation and value at a given site. Multiple sites can be included in the models and visitation and value summed across sites to reflect an entire region. Gravity models work in the opposite direction: total visitation

for an entire region is first estimated, followed by use of the gravity concept where the total visitation is then allocated across sites based on relative attractiveness (Platt, 2000). The aggregate gravity model concept is perhaps more similar to the random utility allocation models under the individual TCM model, while in this paper we focus on single-site TCM studies.

When implementing the TCM in practice, data on visitors' behaviour are often collected on site, because, for sites frequented by only a small proportion of the general population, on-site sampling is much more cost-effective. However, it can lead to problems of endogenous stratification, because frequent users (or, in some cases, visitors who stay longer at the site) will be overrepresented in the sample (Shaw, 1988; Englin & Shonkwiler, 1995). Welfare measures based on the analysis of on-site samples will overstate the benefit derived from access to recreational site, unless the bias in the estimates is corrected. In on-site samples the dependent variable (visits to the site) is truncated at zero, because non-visitors are not observed, and often exhibits overdispersion (the variance of the visits variable exceeds the mean).

Shaw (1998) proposed a correction method for endogenous stratification, applied first to real data by Englin and Shonkwiler (1995). The correction proposed turns out to be very straightforward for equidispersed data which can be assumed to follow a Poisson distribution and has been frequently applied (e.g. Loomis 2003; Hagerty & Moeltner, 2005). However, appropriately correcting for endogenous stratification under overdispersion used to require custom programming by the software user. Only recently, Hilbe and Martínez-Espinoira (2005) packaged the *NBSTRAT* routine, applied to the analyses in this paper, to greatly facilitate this type of analysis using *STATA* (Statacorp, 2005). Achieving convergence is still much more difficult than in the Poisson case, where one simply needs to subtract 1 from the trip count and run a plain Poisson (see Shaw, 1988; or Haab & McConnell, 2002, 174-181, for details). Overdispersion is quite common, because typically the dependent variable takes a low

value in many observations (for example many visitors make few trips to the site or stay only a few days) while it takes a high value for only a few observations (for example, only a few visitors make many trips or stay many days). This means that the variance of the dependent variable in the trip demand function is larger than the mean, making the Poisson model and its variants overly restrictive. Englin and Shonkwiler (1995), Ovaskainen, Mikkola and Pouta (2001), McKean, Johnson and Taylor (2005); and Martínez-Espinoira and Amoako-Tuffour (2008) constitute examples of the few papers where both corrections were applied simultaneously.

By contrasting the effect on welfare measures of applying the correction procedure to different datasets we try in this paper to detect patterns and to identify causal factors behind substantial biases due to on-site sampling.

In the following sections we describe the comparison of estimates corrected for overdispersion, endogenous stratification, and truncation in several recreational demand datasets previously analyzed in the literature. These reanalyses show that zero-truncation accounts for most of the on-site sample bias, as is usually the case (Martínez-Espinoira, Amoako-Tuffour & Hilbe, 2006) but the effect of correcting for endogenous stratification is nevertheless significant. The effect of endogenous stratification is, as the theory predicts, to exaggerate the value of access to a recreational site. However, the relative magnitude of the bias differs depending on the characteristics of the study. In some datasets the effect of accounting for heterogeneous overdispersion is also significant. Furthermore, we find that *NBSTRAT* and *GNBSTRAT* perform the downward correction of welfare estimates and improve goodness of fit only in those cases where we theoretically expect there to be endogenous stratification. Therefore, they can be used not only as a correction mechanism, but also as a diagnosis tool for this bias.

Background

Many travel cost method studies are based on on-site sampling (Englin & Shonkwiler, 1995; Ovaskainen, Mikkola & Pouta, 2001; Bhat, 2003; Shaw, Fadali & Lupi.,

2003; Loomis, 2003, McKean et al., 2003; McKean, Johnson, Taylor & Johnson, 2005). Many recreational activities often attract only a small proportion of the population and users are rarely listed anywhere, so drawing a random sample is very costly. However, as described in further detail by Martínez-Espíñeira et al. (2006), this sampling strategy, which can be seen as a particular type of choice-based sampling, can lead to endogenous stratification. Uncorrected estimates will erroneously overstate the benefit derived from a certain site.

Shaw (1988) considered a correction for endogenous stratification count data estimators in the context of a single recreational site, and derived a correction procedure for the Poisson. Englin and Shonkwiler (1995) extended this correction procedure to the negative binomial model and applied it to real data.

Loomis (2003) compared benefit estimates calculated from a household survey data and data collected on-site, in order to measure the effect of correcting the on-site sample estimates for endogenous stratification. He showed that accounting for the truncated and endogenously stratified nature of the data collected on-site substantially reduced consumer surplus estimates, as theoretically expected, and brought them very close to those unbiased estimates obtained from the household survey. Martínez-Espíñeira et al. (2006) showed in their reanalysis that most of the bias in the uncorrected estimates obtained from Loomis (2003)'s on-site sample was due to the truncation, not the endogenous stratification. Both studies assumed equidispersion in the dependent variable as required by the Poisson. However, Martínez-Espíñeira, Loomis, Amaoko-Tuffour, and Hilbe (2008) reanalyzed the dataset accounting also for overdispersion (with an adjusted negative binomial model) and confirmed the main insights offered by previous comparative analyses.

Apart from those cited above, other studies, such as Ovaskainen et al. (2001), McKean et al. (2003), and McKean et al. (2005) analyzed on-site samples accounting for both overdispersion and endogenous stratification. However, with the exception of Englin and Shonkwiler (1995) and Martínez-Espíñeira and Amoako-Tuffour (2008), previous analyses

apply a negative binomial model that assumes a constant overdispersion parameter for all visitors, while McKean, Johnson, and Taylor (2003) parameterise the overdispersion parameter on an artificially generated variable only. The assumption that the overdispersion parameter is constant across observations is often violated. In the case of conventional count data samples, this prompts the use of a generalized, or heterogeneous (see Hilbe, 2007), negative binomial model that account for this extra source of heterogeneity. This strategy provides information about which predictors contribute to overdispersion, which can be useful when evaluating the model and attempting to determine the worth of each predictor to the model.

For truncated and endogenously stratified data samples, *STATA* 9.1 (Statacorp, 2005) routine *GNBSTRAT* (Hilbe, 2005) can be applied to allow the parameterisation of the overdispersion parameter as a function of visitor characteristics. *NBSTRAT* (Hilbe & Martínez-Espíñeira, 2005) simply optimizes the value of a common overdispersion parameter.

Data

In this article, some analyses are replicated based on household and (mainly) on-site samples available in the literature and extend them to include corrections for overdispersion and/or heterogeneous overdispersion. The reader is referred to the individual source for details about the individual data sets and the results of the analyses conducted in each paper. Here, we will focus on the nature of the dependent variable and the fashion in which the data were collected. We introduce the different datasets in chronological order, based on publication dates.

McConnell et al. (1986)

This dataset, also in Haab and McConnell (2002, pages 156-171) dealt with recreational trips to Fort Phoenix Beach (New Bedford, Massachusetts). There were originally 499 observations in this dataset on five variables, including the round-trip travel costs plus monetary value of time to Fort Phoenix Beach. The data were collected through a household survey, so they contain many zero

values for the dependent variable, but we only use the 168 observations for which the number of trips equals at least one, in order to artificially truncate the sample.

Gurmu and Trivedi (1996)

This is a count data set originally used by Gurmu and Trivedi (1996) with a focus on the modelling of excess zeros, for recreational boating trips collected through a household survey. Discussion of the variables can be found in Sellar, Stoll and Chavas (1985) and Ozuna and Gomaz (1995). These data are also used in Chapter 6.4-6.5 and Chapter 12.6 of Cameron and Trivedi (1998). The dataset includes 659 observations on the number of boating trips to Lake Somerville, East Texas, in 1980 and a series of variables that includes the travel cost to the Lake Somerville, income, and travel costs to substitute lakes. These data were collected through a household survey, so they contain many zero values for the dependent variable. We artificially truncated the sample by restricting our analysis to those 242 observations for which the dependent variable is at least equal to one.

Sohngen, Lichtkoppler and Bielen (2000)

These data were collected to study the value of day trips to Lake Erie beaches. Subsamples of this dataset have also been used by Alberini and Reppas (2005) and Parsons (2003). We use the 223-observation subset (on visits to Maumee Bay State Park beach in Ohio) used by Haab and McConnell (2002 pp. 179-180). This simplified dataset contains only four variables, including number of trips and round-trip travel costs plus monetary value of travel time to that site, round-trip travel costs plus monetary value of travel to nearest substitute beach, and household income.

Ovaskainen et al. (2001)

This dataset contains 656 observations from an on-site survey of visitors conducted on several adjacent recreation sites in the Nuukio Lake Plain, Finland. The dependent variable is the count of trips taken to the site during the last 12 months. A potential anomaly, however, results from the way in which the relevant question was asked ("How many times did you visit this site during the last year?"). Because it

did not explicitly specify whether the current trip should be included or not, there are a non-negligible amount of zeros. This suggests that respondents excluded the current trip, so one trip was added by Ovaskainen et al. (2001) to all observations below 20 trips.

Shrestha, Seidl and Moraes (2002)

Data on recreational fishing were collected from the Brazilian Pantanal over the 3-month period from August through November, 1994. Visitors were queried about their travel costs of the visit, reasons for choosing the Pantanal as a travel destination, aspects of their experiences, and some demographic information. The number of trips taken within the previous year is the dependent variable and it is regressed on several explanatory variables that include the round trip travel cost of the respondent for the current fishing trip.

Loomis (2003)

These data, also used by Martínez-Espíñeira et al (2006) and Martínez-Espíñeira et al. (2006), consist of two sets: one collected through an on-site intercept survey of visitors to the Snake River in Jackson Hole, Wyoming, and a second one collected through a household survey about visits to that same site. Details about the data and the collection process are available in Loomis (2003).

McKean et al. (2003)

McKean et al. (2003) conducted an on-site survey of flat water recreationists on reservoirs at the impounded lower Snake River. The variables used include information on available free time and income, monetary and time costs of travel, outdoor recreation, and other activities on and off the recreation area. The dependent variable is annual trips to the site. McKean et al. (2003) apply a truncated negative binomial regression with an adjustment for endogenous stratification that allows the overdispersion parameter to vary across observations as a function of a randomly generated value. In the appendix they transcribed the code for LIMDEP 7 (Greene, 1995) used to obtain the truncated negative binomial model adjusted for endogenous

stratification and describe their difficulty to achieve convergence with this approach.

Martínez-Espíñeira and Amoako-Tuffour (2008)

This is a subset ($N=413$) of a larger dataset collected on-site at Gros Morne National Park in Newfoundland (Canada). The product of the number of trips to the park in the previous five years times the number of people travelling together in the current trip is regressed against several explanatory variables, including the combined travel cost (money costs and the value of travel time) spent reaching the park and days spent on-site during the current trip. The data were not collected randomly. Visitors were oversampled from rare origins, so the analysis uses sampling weights to correct for this. However, no correction was possible for oversampling of visitors who stayed longer at the park or who visited more locations within the park (so they would have a higher likelihood of being interviewed).

Mendes and Proença (2005)

This is an on-site survey at the Peneda-Gerês National Park (Portugal). The dependent variable is not the number of visits, but rather the number of days on-site during the current visit. In this case, a concern would be the problem of oversampled visitors who stayed longer at the park, since interviewers intercepting visitors within the park would be more likely to find visitors whose visit was longer (a problem described in detail by Lucas, 1963). Crucially, the authors note that, in order to avoid this type of endogenous stratification, visitors were instead interrogated only at the time they addressed themselves to the camping reception centre for camping inscription. For this reason, their reported results do not include a model that corrects for endogenous stratification. The price variable is the on-site and travel out-of-pocket costs, as well as travel and on-site time opportunity costs, and not only travel costs.

Methodology

Count data models are now routinely applied in single-site recreation demand models (Creel & Loomis, 1990; Englin & Shonkwiler, 1995;

Gurmu & Trivedi, 1996; Shrestha et al., 2002). These models account for the fact that the dependent takes only nonnegative integer values. These distributions exhibit a concentration of values on a few, small discrete values (e.g., 0 – 2), skewness to the left, and intrinsic heteroskedasticity with variance increasing with the mean (Cameron & Trivedi, 1998 and 2001). Hellerstein and Mendelsohn (1993) theoretically justified the use of count data to model recreational demand: on any choice occasion, the decision to take a trip is modelled with a binomial distribution. As the number of choices increases the binomial asymptotically converges to a Poisson distribution. The first two moments of the Poisson distribution equal each other, a property known as equidispersion. The model can be extended to a regression framework by parameterizing the relation between the mean parameter and a set of regressors using an exponential mean parametrization.

Overdispersion

However, data on the number of trips are often overdispersed, making the Poisson model overly restrictive. The Poisson maximum likelihood estimator with overdispersion is still consistent, but it underestimates the standard errors and inflates the t-statistics in the usual maximum-likelihood output. If the overdispersion problem is severe, the negative binomial model should be applied. This is commonly obtained by adding an additional parameter that reflects the unobserved heterogeneity that the Poisson fails to capture. This parameter (usually denoted α) determines the degree of dispersion in the predictions (see e. g. Cameron and Trivedi, 1990; Cameron and Trivedi, 2001, p. 336).

Truncation

In on-site samples, the distribution of the dependent variable is also truncated at zero. Ignoring this leads to biased and inconsistent estimates, because the conditional mean is misspecified (Shaw, 1988; Creel & Loomis, 1990; Grogger & Carson, 1991; Yen & Adamowicz, 1993; Englin & Shonkwiler, 1995). In that case, the truncated negative binomial is in order. Examples of applications of this model

include Bowker, English and Donovan (1996); Liston-Heyes and Heyes (1999); and Shrestha et al. (2002). Yen and Adamowicz (1993) compare welfare measures obtained from truncated and untruncated regressions.

Endogenous stratification

Finally, on-site data are affected by endogenous stratification, because a visitors' likelihood of being sampled is positively related to the number of trips they made to the site (or the number of days they spent at the site). If the assumption of equidispersion holds, standard regression packages can be used to estimate a Poisson model adjusted for both truncation and endogenous stratification, as shown by Shaw (1988), by simply running a plain Poisson regression on the dependent variable modified by subtracting 1 from each of its values (Haab & McConnell, 2002, p. 174-181). This model has been used in several applied studies under the assumption of no significant overdispersion (Fix & Loomis, 1997; Hesseln et al., 2003; Loomis, 2003; Hagerty & Moeltner, 2005; Martínez Espiñeira et al., 2006).

For the case where overdispersion is significant, the density of the negative binomial distribution truncated at zero and adjusted for endogenous stratification, derived by Englin and Shonkwiler (1995), cannot be rearranged into an easily estimable form, so it used to require custom programming as a maximum likelihood routine, with the associated increase in computational burden. Englin and Shonkwiler (1995) provide an empirical application of this specification. Englin, Holmes and Sills (2003) and Ovaskainen et al. (2001) also used this model and found that correcting for endogenous stratification on top of zero-truncation does not make much difference in estimates.

However, these studies are based restrict the overdispersion parameter to a common value for all observations (so $\alpha_i = \alpha$). To our knowledge, only Englin and Shonkwiler (1995) have attempted to parameterize α (as $\alpha_i = \alpha 0/\lambda_i$). Ovaskainen, Mikkola, and Pouta (2001) also tried this specification but their keeping α constant at a value previously estimated using a nonlinear squares regression yielded better results in their study. McKean, Johnson, and Taylor (2003) allowed α to vary as a function of

a randomly generated parameter, not related to visitor characteristics. One of the main methodological contributions of the present paper is to use the more flexible approach that allows the overdispersion parameter to vary according to visitor characteristics and compare it with the more restrictive approach. The code is now available for *STATA* 9.1 (Statacorp, 2005) as downloadable commands *NBSTRAT* (Hilbe & Martínez-Espiñeira 2005) and *GNBSTRAT* (Hilbe, 2005). *GNBSTRAT* makes it possible to evaluate how visitors characteristics influence the individual degree of overdispersion and permit to more fully evaluate the effect of these characteristics on the number of trips in the main part of the trip prediction model.

Results

Replicated analyses and the reanalyses of the datasets described in Section Data are considered. In order to check consistency, for all the datasets replicated exactly the analyses conducted in the original works first. Then we ran a negative binomial (NBREG), zero-truncated negative binomial (ZTNB), a zero-truncated negative binomial adjusted for endogenous stratification (NBSTRAT), and a zero-truncated negative binomial adjusted for endogenous stratification and heterogenous overdispersion (GNBSTRAT). These four types of regression are reported in Table 1, summarising the characteristics of the datasets and the results concerning the travel cost coefficient. To maintain consistency, the same model specifications proposed by the original authors to run NBSTRAT and GNBSTRAT is used. For ease of comparison with the original works, the same number of significant decimal places is used to report results.

The focus is on the usefulness of using NBSTRAT and GNBSTRAT and their effects on welfare estimates obtained through count data models. We, therefore assume that the data collection processes and the specifications proposed by the original authors to model the number of trips are a sufficiently valid approximation to the requirements of the individual TCM. In this sense, we abstract, among others, from any potential problems related with additional sources of non-

randomness in the sample (although the idea of oversampling of visitors according to length of stay below is considered) or the fact that some of the datasets might be affected by problems of multi-purpose or multi-site visitation. It is likely that one or more of these internal problems other than those related to the issue of endogenous stratification affect one or more of the studies described below. Those issues are beyond the scope of this work, but the interested reader is directed to Parsons (2003) or Phaneuf and Smith (2006).

McConnell et al (1986)

Using McConnell et al. (1986)'s household sample of beach recreationists, we replicated the Poisson and Negative Binomial specifications reported by Haab and McConnell (2002), not reported, but available upon request, and then applied a zero-truncated model to the positive trip observations of the data set (ZTNB in Table 1). This is compared with the NBSTRAT specification, which not only takes into account truncation and overdispersion, but also endogenous stratification, which should not be expected to affect this dataset.

As expected, NBSTRAT correctly suggests that there is no problem with endogenous in this case, because the data were not collected on site. NBSTRAT yields a worse goodness of fit (log-likelihood) than ZTNB and also a smaller (in absolute value) estimate for the price coefficient, so the consumer surplus per trip, as shown in Table 2, would be higher (\$5.32 while under ZTNB it would be \$5.13). The standard negative binomial regression (NBREG) is also reported, which reveals that correcting for zero-truncation, even in the artificially truncated sample, would account for most of the correction over an inflated estimate of consumer surplus.

Gurmu and Trivedi (1996)

This is a count data set on recreational boating trips to Lake Somerville collected through a household survey. Gurmu and Trivedi (1996) focus on modelling excess zeros. As pointed out by Phaneuf and Smith (2006, p.57) the Poisson and negative binomial distributions typically do not place enough probability mass at zero to match the excess zeros found in many

recreation datasets. Hurdle models consider different data generating processes for explaining the likelihood of individuals being users and for the number of trips for those who are users. There are several types of hurdle and zero-inflated models (see Mullahy, 1986; Lambert, 1992; Cameron & Trivedi, 1998, pp. 123-125 for theoretical details, pp. 889-891; and Martínez-Espíñeira, 2007, for a recent application) and Gurmu and Trivedi (1996) report, among others, the results of a zero-truncated negative binomial as part of their hurdle model. They label this regression Negbin hurdle on Positives.

By restricting the current analysis to the positive values of the dependent variable, we managed to replicate this regression as ZTNB, reported in Table 1, together with the results for NBSTRAT on the positive values of the dependent variable. As expected, this model does not work well on this sample. There was substantial difficulty getting NBSTRAT to converge. Additionally, the log-likelihood worsens relative to ZTNB and the absolute value of the own travel cost coefficient is smaller under NBSTRAT, leading to a higher estimate of consumer surplus per trip (while a correction for endogenous stratification would adjust the consumer surplus downwards).

The command NBSTRAT performs in a satisfactory manner in this example, since even if the researcher had wrongly expected endogenous stratification to affect this household sample, NBSTRAT would have revealed ZTNB preferable to NBSTRAT. Of course the original sample also contains zeros, so the best models overall are either a negative binomial (with no truncation) or, as shown by Gurmu and Trivedi (1996), models that account for excess zeros. We tried to run GNBSTRAT but no choice of independent variables helped explain any additional variation of α across visitors, stressing the notion that, as expected, endogenous stratification is not a problem, so modelling the overdispersion more flexibly while accounting for the nonexistent endogenous stratification was not helpful either.

ON-SITE SAMPLING BIASES & HETEROGENEOUS TRIP OVERDISPERSION

Table 1. Results

Dataset		NBREG	ZTNB	NBSTRAT	GNBSTRAT
McConnell et al. (1986)	β_{TC}	-0.1666**	-0.1950*	-0.1880**	-0.2123**
N = 168 (trips>0 only)	LL	-578.8	-563.3	-564.8	-562.6
Household survey	AIC	1170	1139	1140	1141
Gurmu & Trivedi (1996)	β_{TC}	-0.054***	-0.078***	-0.072***	
N = 242 (trips>0 only)	LL	-644.9	-591.6	-594.3	
Household survey	AIC	1308	1201	1207	
Sohngen et al. (2000)	β_{TC}	-0.013***	-0.017***	-0.017***	-0.029***
N= 223	LL	-588.2	-562.2	-562.3	-549.5
On-site	AIC	1186	1134	1135	1111
Ovaskainen et al. (2001)	β_{TC}	-0.0117***	-0.01484***	-0.01397***	-0.01385***
N= 656	LL	-1928	-1822	-1835	-1834
On-site	AIC	3872	3659	3686	3689
Ovaskainen et al. (2001)	β_{TC}	-0.0098***	-0.01122***	-0.01137***	-0.01095***
N= 542 (trips>1 only)	LL	-1663.8	-1623	-1618	-1611
On-site	AIC	3344	3261	3253	3244
Shrestha et al. (2002)	β_{TC}	-0.0008**	-0.0019***	-0.0021***	-0.0018**
N = 286	LL	-354.5	-175.2	-175.1	-172.4
On-site	AIC	733.1	376.4	376.2	372.8
Loomis (2003)	β_{TC}	-0.02097***	-0.03874***	-0.04076***	-0.02598***
N = 172	LL	-674.5	-624.1	-626.4	-563.3
On-site	AIC	1365	1264	1269	1147
Loomis (2003)	β_{TC}	-0.04617***	-0.06987***	-0.06663***	
N=217	LL	-819.2	-774	-787.8	
Household survey	AIC	1654	1564	1592	
McKean et al. (2003)	β_{TC}	-3.342***	-3.368***	-3.405***	-2.276***
N= 388	LL	-1092.6	-994.4	-995.2	-916.4
On-site	AIC	2213	2017	2018	1865
Martínez-Espiñeira & Amoako-Tuffour (2008)	β_{TC}	-0.3855***	-0.5272***	-0.5701***	-0.4665***
N= 413 (persontrip)	LL	-1020.7	-969.0	-957.6	-940.6
On-site	AIC	2063	1960	1937	1907
Martínez-Espiñeira & Amoako-Tuffour (2008)	$\beta_{Cost/day}$	-0.5709***	-0.7762***	-0.9026***	-0.9051***
N= 413 (days spent on site)	LL	-947.3	-922.6	-908.8	-905.7
On-site	AIC	1915	1865	1838	1833
Mendes & Proença (2005)	$\beta_{Cost/day}$	-0.00526***	-0.00599***	-0.00666***	-0.00614***
N= 243 (days spent on site)	LL	-598.7	-589.5	-590.2	-582.3
On-site	AIC	1211	1193	1194	1185

*p<0.1; **p<0.05; *** p<.001; LL = log-likelihood; AIC = Akaike Information Criterion

Table 2. Consumer surplus estimates.

Dataset		NBREG	ZTNB	NBSTRAT	GNBSTRAT
McConnell et al. (1986)	β_{TC}	-0.1666**	-0.1950*	-0.1880**	-0.2123**
Household survey (trips>0)	CS/trip	\$6.00	\$5.13	\$5.32	\$4.71
Gurmu & Trivedi (1996)	β_{TC}	-0.054***	-0.078***	-0.072***	
Household survey (trips>0)	CS/trip	\$18.51	\$12.90	\$13.88	
Sohngen et al. (2000)	β_{TC}	-0.013***	-0.017***	-0.017***	-0.029***
On-site	CS/trip	\$79.51	\$59.03	\$57.80	\$34.12
Ovaskainen et al. (2001)	β_{TC}	-0.0117***	-0.01484***	-0.01397***	-0.01385***
On-site	CS/trip	\$85.59	\$67.38	\$71.58	\$72.20
Ovaskainen et al. (2001)	β_{TC}	-0.0098***	-0.01122***	-0.01137***	-0.01095***
On-site (trips>1 only)	CS/trip	\$101.75	\$89.12	\$87.94	\$91.32
Shrestha et al. (2002)	β_{TC}	-0.0008**	-0.0019***	-0.0021***	-0.0018**
On-site	CS/trip	\$1250.00	\$526.32	\$476.19	\$555.56
Loomis (2003)	β_{TC}	-0.02097***	-0.03874***	-0.04076***	-0.02598***
On-site	CS/trip	47.68	25.81	24.53	38.49
Loomis (2003)	β_{TC}	-0.04617***	-0.06987***	-0.06663***	
Household survey	CS/trip	\$21.66	\$14.31	\$15.01	
McKean et al. (2003)	β_{TC}	-3.342***	-3.368***	-3.405***	-2.276***
On-site	CS/trip	\$29.93	\$29.69	\$29.37	\$43.94
Martínez-Espiñeira & Amoako-Tuffour (2008)	β_{TC}	-0.3855***	-0.5272***	-0.5701***	-0.4665***
On-site	CS/trip	\$2,593	\$1,897	\$1,754	\$2,143
Martínez-Espiñeira & Amoako-Tuffour (2008)	$\beta_{Cost/day}$	-0.5709***	-0.7762***	-0.9026***	-0.9051***
On-site	CS/day	\$1,752	\$1,288	\$1,108	\$1,105
Mendes & Proença (2005)	$\beta_{Cost/day}$	-0.00526	-0.00599***	-0.00666***	-0.00614***
On-site	CS/day	\$190.11	\$166.94	\$150.15	\$162.87

*p<0.1; ** p<0.05; *** p<.001

Sohngen, et al. (2000)

A subset (N=223) of the original sample was used to successfully replicate the regressions reported by Haab and McConnell (2002, p. 180), who ran a Truncated Poisson model and a Truncated Poisson corrected for endogenous stratification. We report in Table 1 our ZTNB, NBSTRAT and GNBSTRAT results. As expected, endogenous stratification affects the dependent variable in this sample collected on-site. NBSTRAT, although the level of accuracy (3 decimal places) used for the coefficients by Haab and McConnell would not make it apparent, corrects downwards the estimated consumer surplus. Finally, GNBSTRAT was used to model the overdispersion parameter as a function of the travel cost to the site, finding that it significantly improves the goodness of fit. In this case, accounting for the heterogeneous nature of the overdispersion across visitors increases the value of the estimated consumer surplus to \$34.48.

Ovaskainen et al. (2001)

Ovaskainen et al. (2001) reported the results of running in LIMDEP (Greene, 1995) a series of count data models that include zero-truncated models and also models that correct for endogenous stratification. The replication with NBSTRAT in STATA yields slightly different results than what the original authors report as their zero-truncated endogenously stratified negative binomial. This is likely due to the fact that they had to fix the value of alpha to a constant estimated from a separate regression based on nonlinear least squares. It is noteworthy that NBSTRAT achieved a much higher log-likelihood (-1835) than the original procedure used by the original authors (-1891). Additionally, note that the zero-truncated endogenously stratified negative binomial yields a price coefficient (-0.01397) that is smaller in absolute value than the one obtained without correcting for endogenous stratification. This is in line with the results obtained in Ovaskainen et al. (2001). Because the data were collected on-site, we would expect a bias from endogenous stratification in the opposite direction.

It is possible that this puzzling result has to do with the anomaly in the data described in Section Data. The original authors added one

trip to each observation with trips less than 20, being unsure of whether respondents had included the current trip in their response or not. This possibility seems more likely when we analyze only those 541 observations for which the 'manipulated' number of trips is more than one. If that is done NBSTRAT performs as expected. Although not reported here, further regressions on smaller samples (for observations with only more than 2 trips, more than 3 trips, etc) confirmed in an increasingly reassuring way that the endogenous stratification correction performed by NBSTRAT would have worked in the expected direction if the data collection had not suffered from this unfortunate wording of the question about the number of trips.

It can also be shown that, for the trimmed samples, GNBSTRAT also slightly overperforms the previous models by making the overdispersion parameter a function of the age, equipment ownership, and income of the visitors. GNBSTRAT results are reported, although they do not offer much improvement over NBSTRAT.

Shrestha, et al. (2002)

A model equivalent to ZTNB was reported in the original paper. We failed to replicate its results exactly, but they are similar. The original authors claimed that no significant bias due to endogenous stratification was expected, "mainly because of the one-time survey of the anglers, rather than using annual visitor-data in the analysis." NBSTRAT shows (Table 1) that the correction would clearly reduce the estimates of consumer surplus per trip (from \$526.32 to \$476.1). However, the improvement in terms of log-likelihood is not substantial. GNBSTRAT improves the fit somewhat by making the overdispersion parameter a function of income.

Loomis (2003)

Income was rescaled into \$10,000 units, but otherwise the same 172 observations and variables were used when applying different count data specifications to the on-site sample used in Loomis (2003). NBSTRAT performs the appropriate type of correction on ZTNB. However, NBSTRAT does not improve the fit much. A GNBSTRAT specification that makes

the overdispersion parameter α a function of the number of trips and income does improve the fit substantially.

When it comes to the reanalysis of the household sample collected by Loomis (2003), it can be seen in Table 1 that NBSTRAT would, as expected, show no improvement over ZTNB on an artificially truncated sample. The log-likelihood worsens and the estimated consumer surplus per trip increases, while a correction for endogenous stratification on a sample collected on-site would of course lead to a measure of consumer surplus revised downwards.

McKean et al (2003)

Using the code provided by McKean et al. (2003) in an appendix, the results were replicated using the maximizing commands in LIMDEP (Greene, 1995). However, when GNBSTRAT in STATA was used to try to replicate them, we failed to obtain the same results. This is probably because of two reasons. First, McKean et al. (2003) parameterise their overdispersion parameter as a function of a randomly generated variable (zz in their own notation) which takes a different value in each estimation. Second, it is likely that STATA maximum likelihood routine can obtain a more finely improved log-likelihood.

In any case, the results are similar in regards to the correction for endogenous stratification. In Table 1 we show the results of several specifications using the same data set used in the original. However, note that the values were rescaled (dividing by 100) for some of the variables to improve the presentation. For example the estimate found by McKean et al. (2003) for the travel cost under the equivalent of NBSTRAT was equal to -0.0337, while we obtained -3.405. The goodness of fit improves as we allow for a more flexible specification that accounts for on-site sample biases. NBSTRAT performs the expected type of correction on the estimates of consumer surplus per trip. In this case the magnitude of the bias caused by on-site sampling is not substantial in terms of consumer surplus per trip.

One off-pattern feature of the analysis of this dataset is that the correction for zero-truncation in itself does not seem to account in this case for much of the correction of the bias

due on-site sampling. As suspected, this appears to be related to the high value of the average number of trips (8.448). This value is larger than in the other studies analyzed, but closest to the equivalent value in Ovaskainen et al. (2001). The results from the latter show that, most of all when only trip values above one are used, the zero-truncation correction is, by itself, not substantial in relative terms. This is in line with the intuition that this type of correction is more necessary when the average value of the count (trips in the illustrations used here) is very low, as is typical in count data analyses.

Martínez-Espíñeira and Amoako-Tuffour (2008)

Martínez-Espíñeira and Amoako-Tuffour (2008) considered the dependent variable *persontrip* (number of trips times size of visitor party) as a function of travel costs and other characteristics of the trip and the visitors. Here we use a subsample of their data set to illustrate the effect of correcting for endogenous stratification. As expected, NBSTRAT performs a downward correction on the estimates of consumer surplus per trip. Allowing the overdispersion parameter to vary according to variables related to income and the age composition of the visitor party would improve the goodness of fit.

With this dataset it can also be considered how the issue of on-site sampling can affect welfare estimates when the travel cost model is based on the length of stay as the dependent variable (Lucas, 1963; Mendes & Proença 2005). In this case visitors were intercepted at several locations within the park, with no specific strategy for avoiding oversampling those visitors who stayed longer at the park. Therefore, those visitors who spend more days at the park had a higher likelihood of being intercepted than those who spent fewer days. Correcting for the resulting endogenous stratification would reduce the estimates of consumer surplus in a model that relates the length of stay to the cost of reaching the park. For this reanalysis, combined travel and on-site cost per day was constructed analogous to the one used by Mendes and Proença (2005). The associated coefficient are labeled $\beta_{\text{Cost/day}}$ in Tables 1 and 2. Those visitors who face a higher combined travel and stay cost are expected to

spend less time at the site. They likely use part of their available recreational time to visit other sites adjacent to the site of interest or on the way to it from their home. Correcting for endogenous stratification in this case also works as expected, decreasing the estimate of consumer surplus per day spent at the park.

It is noteworthy that in this case, the number of trips made to the park has no significant effect on the length of stay during the current trip. Only when GNBSTRAT makes the overdispersion parameter a function of that variable does the number of trips become significant and does it take the expected negative sign. It was expected to find that those who live closer to the park make more frequent but shorter visits to the park, once every other influence on the length of stay (particularly the travel cost) has been controlled for.

Mendes and Proença (2005)

Contrary to the case of Martínez-Espiñeira and Amoako-Tuffour (2008), Mendes and Proença (2005) did adopt a specific strategy to avoid oversampling those visitors who stayed longer at the park. They only interviewed them when signing in at the camping reception centre. This strategy is expected to successfully avoid the problem of endogenous stratification, so NBSTRAT was used to test if this was indeed true.

Table 1 shows that, although the estimated coefficient of the price variable (minimum recreation cost of each day of stay at the site, including travel cost) is slightly larger in absolute terms under NBSTRAT than under ZTNB, there is no improvement in goodness of fit due to correcting for endogenous stratification under the negative binomial models. Once again, this confirms that NBSTRAT can be relied upon to diagnose problems of endogenous stratification, since it does not spuriously improve the goodness of fit, relative to the uncorrected ZTNB for samples that are not affected by the problem.

Conclusion

The reanalyses above show that the newly developed commands NBSTRAT and GNBSTRAT perform appropriately when correcting for the simultaneous problems of

zero-truncation, overdispersion, and endogenous stratification. These commands illustrate the effect of endogenous stratification on the estimates obtained from samples of recreation data obtained on site and allow the researcher to easily correct the resulting bias. Furthermore, by applying them to datasets obtained by artificially truncating at zero a sample collected from the general population, we show that the commands will not reduce the estimates of consumer surplus and will not improve the goodness of fit of the regression when they are applied to datasets that are not actually affected by endogenous stratification. That is, it is not a sledgehammer solution: it only works well when the problem is actually there. In this sense, we can safely suggest the use of NBSTRAT and GNBSTRAT as both a diagnostic tool, useful when the researcher does not know how serious the problem of oversampling of avid users is, and as a correction tool for the bias. NBSTRAT helped us confirm that, in some cases, on-site sampling is not subject to endogenous stratification if the sampling strategy is carefully designed to avoid it.

We have confirmed for several datasets that most of the overall bias caused by sampling on site is due to the truncation at zero of the dependent variable. This is a result that appears to apply regardless of the idiosyncrasies of each particular example, although it is less apparent in those datasets with a high average value of the dependent variable. However, the problem of endogenous stratification contributes to inflate uncorrected welfare estimates.

We expect that these two newly developed commands will help applied researchers with average computing abilities to properly analyze recreational datasets obtained through on-site surveys. By applying them, while enjoying the advantages of on-site sampling, researchers no longer need to worry about endogenous stratification or the computational burden associated to alternative ways to handle it.

Note that in the analysis we have assumed that the only problems affecting the welfare estimates had to do with on-site sampling. In particular, we assumed that the assumptions needed for a meaningful travel cost method analysis were met and that the correct

set of variables was included in the model specification in each case. Further research efforts should be directed at addressing these issues and analysing the influence of different types of misspecification and measurement problems on the magnitude of biases due to on-site sampling.

Finally, we should note that although the corrections showcased in this paper focused on the effects on consumer surplus measures in the context of the travel cost method, the analysis extends to any other type of count data analysis where obtaining unbiased estimates of the relevant coefficients was an issue.

References

- Alberini, A. & D. Reppas (2005). Model misspecification and endogenous on-site sampling in the travel cost method. presented at EAERE 2005 Bremen.
- Bhat, M. G. (2003). Application of non-market valuation to the Florida Keys Marine Reserve Management. *Journal of Environmental Management* 67 (4), 315-325.
- Bowker, J. M., D. B. K. English & J. A. Donovan (1996). Toward a value for guided rafting on southern rivers. *Journal of Agricultural and Applied Economics* 28 (2), 423-432.
- Braden, J. B. & C. Kolstad (1991). *Measuring the Demand for Environmental Quality*. Amsterdam: Elsevier.
- Cameron, A. C. & P. K. Trivedi (1990). Regression-based tests for overdispersion in the poisson model. *Journal of Econometrics* 46 (3), 347-364.
- Cameron, A. C. & P. K. Trivedi (2001). Essentials of count data regression. In B. H. Baltagi (Ed.), *A Companion to Theoretical Econometrics*, pp. 331-348. Oxford, U.K.: Blackwell.
- Cameron, C. & P. K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.
- Creel, M. & J. B. Loomis (1990). Theoretical and empirical advantages of truncated count data estimators for analysis of deer hunting in California. *American Journal of Agricultural Economics* 72, 434-441.
- Englin, J. & J. Shonkwiler (1995). Estimating social welfare using count data models: An application under conditions of endogenous stratification and truncation. *Review of Economics and Statistics* 77, 104-112.
- Englin, J. E., T. P. Holmes & E. O. Sills (2003). Estimating forest recreation demand using count data models. In E. O. Sills (Ed.), *Forests in a Market Economy*, Chapter 19, pp. 341-359. Dordrecht, The Netherlands: Kluwer Academic Publishers.
- Fix, P. & J. Loomis (1997). The economic benefits of mountain biking at one of its meccas: An application of the travel cost method to mountain biking in Moab, Utah. *Journal of Leisure Research* 29 (3), 342-352.
- Freeman III, A. M. (1993). *The Measurement of Environmental and Resource Values: Theory and Methods*. Washington D.C.: Resources for the Future.
- Greene, W. H. (1995). *LIMDEP Version 7.0 User's Manual*. Bellport, NY: Econometric Software, Inc.
- Grogger, J. T. & R. T. Carson (1991). Models for truncated counts. *Journal of Applied Econometrics* 6 (3), 225-238.
- Gurmu, S. & P. Trivedi (1996). Excess zeros in count models for recreational trips. *Journal of Business and Economic Statistics* 14, 469-477.
- Haab, T. & K. McConnell (2002). *Valuing Environmental and Natural Resources: Econometrics of Non-Market Valuation*. Cheltenham, UK: Edward Elgar.
- Hagerty, D. & K. Moeltner (2005). Specification of driving costs in models of recreation demand. *Land Economics* 81 (1), 127-143.
- Hellerstein, D. & R. Mendelsohn (1993). A theoretical foundation for count data models. *American Journal of Agricultural Economics* 75 (3), 604-611.
- Hesseln, H., J. B. Loomis, A. González-Cabán & S. Alexander (2003). Wildfire effects on hiking and biking demand in New Mexico: A travel cost study. *Journal of Environmental Management*, 69 (4), 359-368.

Hilbe, J. M. (2005). GNBSTRAT: Stata module to estimate generalized negative binomial with endogenous stratification. Statistical Software Components, Boston College Dept. of Economics: <http://ideas.repec.org/c/boc/bocode/s456413.html>.

Hilbe, J. M. & R. Martínez-Españeira (2005). NBSTRAT: Stata module to estimate negative binomial with endogenous stratification. Statistical Software Components, Boston College Department of Economics. <http://econpapers.repec.org/software/bocbocode/s456414.htm>.

Hilbe, J.M (2007). *Negative Binomial Regression*, Cambridge: Cambridge University Press.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 34, 1-14.

Liston-Heyes, C. & A. Heyes (1999). Recreational benefits from the Dartmoor National Park. *Journal of Environmental Management* 55 (2), 69-80.

Loomis, J. (2003). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure. *Water Resources Research* 39 (4), 1105.

Lucas, R. C. (1963) Bias in estimating recreationists' length of stay from sample interviews. *Journal of Forestry*, 61(2), 912-913

Martínez-Españeira, R. (2007) "Adopt a Hypothetical Pup': A Count Data Approach to the Valuation of Wildlife" *Environmental and Resource Economics*, 37(2), 335-360.

Martínez-Españeira, R. & J. Amoako-Tuffour (2008). Recreation Demand Analysis under Truncation, Overdispersion, and Endogenous Stratification: An Application to Gros Morne National Park, *Journal of Environmental Management*, 88(4), 1320-1332

<http://ideas.repec.org/p/wpa/wuwpem/051107.html>.

Martínez-Españeira, R., J. Amoako-Tuffour & J. M. Hilbe (2006a). Travel cost demand model based river recreation benefit estimates with on-site and household surveys: Comparative results and a correction procedure: Reevaluation. *Water Resources Research* 42 (W10418). DOI:10.1029/2005WR004798.

Martínez-Españeira, R., J. B. Loomis, J. Amoako-Tuffour & J. M. Hilbe (2008). Comparing recreation benefits from on-site versus household surveys in count data travel cost demand models with overdispersion, *Tourism Economics* 14(3), 567-576

McConnell, K. E. & I.E.C. Inc. (1986). The damages to recreational activities from PCBs in New Bedford Harbor. Unpublished report prepared for NOAA Ocean Assessment Division. Rockville, MD: December.

McKean, J. R., D. Johnson & R. G. Taylor (2003). Measuring demand for .at water recreation using a Two-Stage/Disequilibrium travel cost model with adjustment for overdispersion and self-selection. *Water Resources Research* 39 (4), 1107.

McKean, J. R., D. Johnson, R. G. Taylor & R. L. Johnson (2005). Willingness to pay for non angler recreation at the Lower Snake River reservoirs. *Journal of Leisure Research* 37 (2), 178-191.

Mendes, I. & I. M. Proença (2005). Estimating the recreation value of ecosystems by using a travel cost method approach. WP 2005/08 DE/CIRIUS, Instituto Superior de Economia e Gestão, Universidade Técnica de Lisboa. <http://ideas.repec.org/p/ise/isegwp/wp82005.html>

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics* 33, 341-365.

Ovaskainen, V., J. Mikkola & E. Pouta (2001). Estimating recreation demand with on-site data: An application of truncated and endogenously stratified count data models. *Journal of Forest Economics* 7 (2), 125-144.

Ozuna, T. & I. A. Gomez (1995). Specification and testing of count data recreation demand functions. *Empirical Economics* 20 (3), 543-550.

Parsons, G. R. (2003). The travel cost model. In P. A. Champ, K. J. Boyle & T. C. Brown (Eds.), *A Primer on Nonmarket Valuation*, Chapter 9. London: Kluwer Academic Publishing.

Phaneuf, D. J. & V. K. Smith (2006). Recreation Demand Models. In K. G. Mäler & J. R. Vincent (Eds.) *Handbook of Environmental Economics*, Chapter 15. North-Holland Elsevier Vol. 2, Number 2.

Platt, J. 2000. "Impact of Fluctuating Reservoir Elevation on Recreation Use and Value." Technical Memorandum #: EC-2000-02, U. S. Bureau of Reclamation, Technical Service Center, Economics Group, Denver CO.

Sellar, C., J. Stoll & J. Chavas (1985). Validation of empirical measures of welfare change: A comparison of nonmarket techniques. *Land Economics* 61, 156-175.

Shaw, D. (1988). On-site sample regression: Problems of non-negative integers, truncation, and endogenous stratification. *Journal of Econometrics* 37, 211-223.

Shaw, W. D., E. Fadali & F. Lupi (2003). Comparing consumer's surplus estimates calculated from intercept and general survey data. Proceedings of the W-133 (U.S.D.A.) Regional Economics Group, compiled by J. Scott Shonkwiler. Las Vegas, Nevada, February.

Shrestha, K. R., A. F. Seidl & A. S. Moraes (2002). Value of recreational fishing in the Brazilian Pantanal: A travel cost analysis using count data models. *Ecological Economics* 42 (1-2), 289-299.

Sohnngen, B., F. Lichtkoppler & M. Bielen (2000). The value of day trips to Lake Erie beaches. Technical Report TB-039, Ohio Sea Grant Extension, Columbus OH.

StataCorp (2005). *Stata Statistical Software: Release 9.1*. College Station, TX: StataCorp LP.

Yen, S. T. & W. L. Adamowicz (1993). Statistical properties of welfare measures from count-data models of recreation demand. *Review of Agricultural Economics* 15, 203-215.

Computing Multivariate Process Capability Indices (Excel)

Michele Scagliarini
University of Bologna

Raffaele Vermiglio
Telecom Italia Mobile

In manufacturing industry there is growing interest in measures of process capability under multivariate setting. Although there are many statistical packages to assess univariate capability, a current problem with the multivariate measures of capability is the shortage of user friendly software. In this article a Visual Basic program has been developed to realize an Excel spreadsheet that may be used to compute two multivariate measures of capability. The aim of this article is to provide a useful tool for practitioners dealing with multivariate capability assessment problems. The features of the program include easy data entry and clear report format.

Key words: Multivariate process capability indices, statistical quality control, Visual Basic, Excel

Introduction

Process capability indices have been widely used in the manufacturing industry providing numerical measures on process performances. Juran et al. (1974) first introduced the idea of capability indices (the original name was capability ratios). The first indices were univariate, measured the process capability with regard to a single quality measure and focused on the percentage of non-conforming Kotz and Lovelace (1998).

In recent years multivariate capability indices were developed as a natural extension to the univariate concept. Multivariate capability indices appeared in the literature during the early 1990s. Most of them assumed multivariate normal data, a stable process, and were generalizations of their univariate counterparts.

Michele Scagliarini is Associate Professor of Statistics. Research interests include statistical methods for quality improvement and environmental statistics. Email: michele.scagliarini@unibo.it. Raffaele Vermiglio's research interests include statistical methods for quality control and statistical software. Email: raffaelevermiglio@tim.it.

Wang et al. (2000) compared three multivariate indices: the multivariate capability vector, Shahriari et al. (1995); the multivariate capability index MC_{pm} , Taam et al. (1993); the multivariate capability index for process potential MC_p , Chen (1994). Although there are many statistical packages for running univariate capability analysis, a current problem with the multivariate measures of capability is the shortage of user friendly software. Recently, an interesting contribution is the work by Phnadsnis et al. (2005). The authors proposed a Visual Basic program to perform bivariate capability analysis using the MC_{pm} index with Excel.

In this article, a set of Visual Basic macros are developed that may be used to compute multivariate measures of capability using Excel. Since Excel is often used by engineers, or generally by non statisticians, our purpose is to provide a user-friendly tool to help practitioners in performing multivariate capability analysis. Two different multivariate capability measures are examined: the multivariate capability vector, Shahriari et al. (1995) and the multivariate capability index MC_{pm} , Taam et al. (1993). The two multivariate capability measures are described, and the software is described.

Two multivariate capability indices

Assume that a process can be described by a v -dimensional vector of measurements $\mathbf{x}$ and we further assume that the joint probability

distribution of the v quality characteristics is the multivariate normal distribution

$$\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (1)$$

The multivariate capability vector

The multivariate capability vector was proposed by Shahriari *et al.* (1995) based on the original work of Hubele *et al.* (1991). The vector consists of three components and appears as

$$(C_{pM}, PV, LI)$$

The first component of the vector, C_{pM} , is a ratio of areas or volumes, analogous to the ratio of lengths of the univariate C_p index. The numerator is the area (two-dimensional case) or the volume (three or more dimensions) defined by the engineering tolerance region, while the denominator is the area or volume of a modified process region, defined as the smallest region similar in shape to the engineering tolerance region, circumscribed about a specified probability contour:

$$C_{pM} = \left[\frac{\text{Volume of engineering tolerance region}}{\text{Volume of modified process region}} \right]^{1/v} \quad (2)$$

where v is the number of characteristics of the process. The volume of the engineering tolerance region is

$$\prod_{i=1}^v (USL_i - LSL_i) \quad (3)$$

where USL_i and LSL_i are the upper and lower limits respectively, relative to the characteristic i ($i=1,2,\dots,v$).

To compute the volume of the modified process region it is worth reminding that under the hypothesis of multivariate normality the statistic

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu}) = g(\mathbf{x}) \quad (4)$$

follows a χ^2 distribution with v degree of freedom. Therefore, the borders of the process region UPL_i , the upper process limit, and LPL_i , the lower process limit ($i=1,2,\dots,v$) are

determined by solving the systems of equations of first derivative, with respect to each x_i of the quadratic form

$$(\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma} (\mathbf{x} - \boldsymbol{\mu}) = \chi^2_{(v,\alpha)} \quad (5)$$

where $\chi^2_{(v,\alpha)}$ is the upper $100(\alpha)$ of a χ^2 distribution with v degrees of freedom associated with the probability contour. Usually, in analogy with the “ 6σ ” in the denominator of the univariate indices, $\alpha=0.0027$.

The solutions (two for each dimension i) of equation (5) are Wang *et al.* (2000):

$$UPL_i = \mu_i + \sqrt{\frac{\chi^2_{(v,\alpha)} \det(\boldsymbol{\Sigma}_i^{-1})}{\det(\boldsymbol{\Sigma}^{-1})}} \quad (6)$$

$$LPL_i = \mu_i - \sqrt{\frac{\chi^2_{(v,\alpha)} \det(\boldsymbol{\Sigma}_i^{-1})}{\det(\boldsymbol{\Sigma}^{-1})}} \quad (7)$$

where $i=1,2,\dots,v$ and $\det(\boldsymbol{\Sigma}_i^{-1})$ is the determinant of a matrix obtained from $\boldsymbol{\Sigma}^{-1}$ by deleting the i^{th} row and column. Thus, the volume of the modified process region is

$$\prod_{i=1}^v (UPL_i - LPL_i) \quad (8)$$

In practice $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and their sample estimates $\bar{\mathbf{x}}$ and $\mathbf{S}$ can be used:

$$\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \quad (9)$$

$$\mathbf{S} = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (10)$$

Values of C_{pM} higher than one indicate that the modified process region is smaller than the engineering tolerance region, therefore we have high probability that the produced items will be classified as conform.

The second component of the vector is defined as the significance level of a Hotelling's T^2 statistic computed under the assumption that the center of the engineering specifications is

COMPUTING MULTIVARIATE PROCESS CAPABILITY INDICES

considered to be the true underlying mean of the process:

$$PV = \Pr \left(T^2 > \frac{v(n-1)}{(n-v)} F_{(v, n-v)} \right) \quad (11)$$

where

$$T^2 = n(\bar{\mathbf{x}} - \boldsymbol{\mu})' \mathbf{S}^{-1}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \quad (12)$$

and $F_{(v, n-v)}$ is the F distribution with v and $n-v$ degrees of freedom.

Values of PV close to zero indicate that the center of the process is far from the engineering target value.

The third component of the vector summarizes a comparison of the location of the modified process region and the tolerance region. It indicates whether any part of the modified process region falls outside the engineering specifications. It has a value of 1 if the entire modified process region is contained within the tolerance region and, otherwise, a value of 0:

$$LI = \begin{cases} 1 & \text{if modified process region is contained} \\ & \text{within the tolerance region} \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

The multivariate capability index MC_{pm}

The index MC_{pm} was proposed by Taam *et al.* (1993) and is defined as a ratio of two volumes. The numerator is the volume of the modified tolerance region R_1 and the denominator is the volume of the scaled 99.73 percent process region R_2 . Under the multinormality hypothesis we have an elliptical process region, while the modified tolerance region is the largest ellipsoid that is centered at the target completely within the original tolerance region.

In the general case of v characteristics R_1 is an hyperellipsoid and the volume is given by Kendall (1961)

$$\text{Vol.}(R_1) = \frac{2 \prod_{i=1}^v a_i}{v} \frac{\pi^{v/2}}{\Gamma(\frac{v}{2})} \quad (14)$$

where the a_i ($i=1,2,\dots,v$) are the lengths of the semi-axes. Then the multivariate capability index is written as

$$MC_{pm} = \frac{\text{Vol.}(R_1)}{\text{Vol.}((\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}_T^{-1} (\mathbf{x} - \boldsymbol{\mu}) \leq K(v))} \quad (15)$$

where $\mathbf{x}$ is the vector ($v \times 1$) of measurements from a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_T = E[(\mathbf{x} - \mathbf{T})(\mathbf{x} - \mathbf{T})']$ is the mean square error matrix from the process, $\mathbf{T}$ is a vector of target values, and $K(v)$ is a 99.73th percentile of a χ^2 with v degrees of freedom.

The denominator of MC_{pm} can be also expressed as a product of two terms:

$$\begin{aligned} \text{Vol.}(R_2) &= |\boldsymbol{\Sigma}|^{1/2} (\pi K(v))^{v/2} [\Gamma(v/2 + 1)]^{-1} \times \\ &\times \left[1 + (\boldsymbol{\mu} - \mathbf{T})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{T}) \right]^{-1/2} = \\ &= \text{Vol.}(R_3) \times \left[1 + (\boldsymbol{\mu} - \mathbf{T})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{T}) \right]^{-1/2} \end{aligned} \quad (16)$$

where R_3 is the region in which 99.73% of the process values fall within.

Therefore MC_{pm} can be rewritten as:

$$MC_{pm} = \frac{\text{Vol.}(R_1)}{\text{Vol.}(R_3)} \frac{1}{\left[1 + (\boldsymbol{\mu} - \mathbf{T})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{T}) \right]^{1/2}} = \frac{C_p}{D} \quad (17)$$

The MC_{pm} index is a function of two components: C_p which represents the process variability relative to the modified tolerance region; D which detects the process deviation from the target. Given a random sample of n measurements, $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ each of dimension v , the estimator for MC_{pm} is given by

$$\hat{M}C_{pm} = \frac{\text{Vol.}(R_1)}{|\mathbf{S}|^{1/2} (\pi K)^{v/2} [\Gamma(v/2+1)]^{-1}} \times \frac{1}{\left[1 + \frac{n}{n-1} (\bar{\mathbf{x}} - \mathbf{T})' \mathbf{S}^{-1} (\bar{\mathbf{x}} - \mathbf{T})\right]^{1/2}} = \frac{\hat{C}_p}{\hat{D}} \quad (18)$$

When the process mean vector equals the target vector, and the index has the value 1, then 99.73% of the process values lie within the modified tolerance region. The numerator C_p is analogous to the univariate C_p , that is, a value greater than 1 implies that the process has smaller variation than allowed by the specification limits with a certain confidence level; a value less than 1 implies more variation. Also, $0 < 1/\hat{D} < 1$ measures the closeness between the process mean and the target; a larger $1/\hat{D}$ indicates that the mean is close to target.

Software Description

The macros are stored in the MultiCap.xls file. The user can directly open this file to perform the capability analysis. To illustrate the software consider a simulated example. We generate a sample of 100 observation from a multivariate normal process of dimension $v=3$ with mean vector and covariance matrix given by

$$\boldsymbol{\mu}' = [40 \quad 60 \quad 15]$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.100 & 0.483 & 0.308 \\ 0.483 & 0.4 & 0.185 \\ 0.308 & 0.185 & 0.600 \end{bmatrix}$$

respectively.

The target values coincide with the means and the specification limits are reported in Table 1. The user interface is the worksheet "INPUT" (Figure 1) where the main parameters of the analysis can be specified. Clicking the button "PARAMETERS" the form "ANALYSIS PARAMETERS" will appear, as shown in Figure 2, thus can be specified: the number v of quality characteristics *i.e.* the dimensions of the process (for this version the maximum number is 5); the α value to define the size of the tolerance region (usually $\alpha = 0.0027$); the sample size. Therefore, these values will be displayed in the Table "GENERAL PARAMETERS" and the Tables "SPECIFICATIONS LIMITS AND TARGET" and "SAMPLE DATA" will be automatically modified on the basis of the values of v and n .

Clicking the button "LIMITS-TARGET" the corresponding form (Figure 3) will appear, thus for each quality characteristic can be entered: the specification limits; the target value. Clicking the button "SAMPLE DATA" the form "INPUT SAMPLE DATA" (Figure 4) will appear, in this way the sampling observations can be inserted. If the sampling observations are already available, then the data can be directly pasted in the Table "SAMPLE DATA".

Using the simulated data the worksheet appears as shown in Figure 5. The calculations can be performed clicking the button "COMPUTE". The procedure is splitted in two separate steps. The first step consists in the estimation of the mean vector and covariance matrix. The results of the computations are displayed in the worksheet "MEANS AND COVARIANCES" as shown in Figure 6.

In the second step the user can choose between the two capability measures (Figure 7). The results will be displayed, together with brief report, in a suitable form (Figures 8 and 9). In the report box are comments on the results. The messages reported here are only examples and can be modified if necessary.

COMPUTING MULTIVARIATE PROCESS CAPABILITY INDICES

Table 1. Specification limits

Characteristic	LSL_i	USL_i
1	33	47
2	52	68
3	12	18

Figure 1. The worksheet INPUT

UserForm1

ANALYSIS PARAMETERS

Number of characteristics (dimensions)

v

Probability value for the natural NC (usually 0.0027)

α

Sample size

n

OK

CANCEL

Figure 2. User form “ANALYSIS PARAMETERS”

UserForm1

LIMITS AND TARGET

CHARACTERISTIC NUMBER

SPECIFICATION LIMITS

LSL

USL

TARGET VALUE

T

OK CANCEL

Figure 3. User form “LIMITS AND TARGET”

DATA INPUT

INPUT SAMPLE DATA

SAMPLE UNIT NUMBER N°

VALUE OF CHARACTERISTIC 1

VALUE OF CHARACTERISTIC 2

OK CANCEL

Figure 4. User form “INPUT SAMPLE DATA”

PARAMETERS AND DATA

GENERAL PARAMETERS

v	3
α	0.0027
n	100

SPECIFICATION LIMITS AND TARGET

CHARACTERISTIC BY	LSL	USL	TI
1	33	47	40
2	52	68	60
3	12	18	15

PARAMETERS

PARAMETERS
LIMITS-TARGET
SAMPLE DATA
COMPUTE
RESET
INSTRUCTIONS

SAMPLE DATA

UNITS	CHARACTERISTICS	1	2	3
1		39.622	60.985	15.06
2		40.861	60.161	13.984
3		39.592	61.384	14.4
4		40.12	57.971	16.783
5		42.347	59.432	13.904
6		39.203	56.966	14.451
7		40.837	61.872	15.956
8		38.394	60.173	13.943
9		40.638	59.333	15.347
10		38.822	58.357	13.661
11		39.642	60.019	15.217
12		39.285	60.706	15.807
13		40.124	60.813	15.592
14		40.035	58.593	14.875
15		39.787	59.225	14.129
16		41.815	60.476	16.135

Figure 5. A portion of the worksheet “INPUT” with the example's data

COMPUTING MULTIVARIATE PROCESS CAPABILITY INDICES

MEANS AND COVARIANCES

MEAN VECTOR

X1	X2	X3
39.91997	60.01997	15.04785

COVARIANCE MATRIX

0.99678398	0.806763	0.27429176
0.806763	1.67344893	0.16861163
0.27429176	0.16861163	0.6560348

det(S)
0.754664509

det(S⁻¹)
1.325092128

S⁻¹

1.417067894	-0.466179151	-0.472668261
-0.466179151	0.766815443	-0.002171995
-0.472668261	-0.002171995	1.722492632

DET(S⁻¹ * 100)

1.320829386	2.217474005	0.868306543
-------------	-------------	-------------

Figure 6. Worksheet MEANS AND COVARIANCES

Figure 7. Index choice

CAPABILITY ANALYSIS

ANALYSIS RESULTS
VECTOR (CpM, PV, LI)

CpM	1.44
PV	0.62
LI	0

REPORT

The process is CAPABLE.

OK

Figure 8. Results and report

CAPABILITY ANALYSIS

ANALYSIS RESULTS
MCpm INDEX

INDEX

MCpm	3.6
------	-----

INDEX COMPONENTS

MCp	3.63
D	1.01

REPORT

THE PROCESS IS CAPABLE. The process is close to the target.

OK

Figure 9. Results and report

COMPUTING MULTIVARIATE PROCESS CAPABILITY INDICES

Figure 10. The upper portion of the worksheet "VECTOR"

Figure 11. The lower portion of the worksheet "VECTOR"

Figure 12. The upper portion of the worksheet “INDEX MC_{pm}”

Figure 13. The lower portion of the worksheet “INDEX MC_{pm}”

RESULTS	
MC_{pm}	3.60
MC_p	3.63
D	1.01
C_{pM}	1.44
PV	0.62
LI	0

Figure 14. Worksheet “OUTPUT”

In the worksheet “VECTOR” (Figure 10 and Figure 11) are reported the details of the calculations for the multivariate capability vector. Similarly, in the worksheet “INDEX MC_{pm} ” (Figure 12 and 13) are reported the details of the calculations for the MC_{pm} index. Results concerning both capability measures are summarized in the worksheet “OUTPUT” as shown in Figure 14.

Some details

The Visual Basic code for the calculation management is stored in the Macro CALCOLO1. The inversion of the matrices and the computation of the determinants are performed using the Excel functions MInverse and MDeterm respectively. Moreover, we used the function INV.CHI() to calculate the quantiles of the chi-square distributions in the computations of the process regions and the function DISTRIB.F() to compute the PV component of the multivariate vector. As

pointed out by Knusel (1998) and McCulloch and Wilson (2002), it is important to take into account the accuracy problems when using these Excel functions. Keeping this caution in mind, in agreement with the purposes of the present work, we retain the degree of precision acceptable.

Conclusion

An Excel spreadsheet is presented which can be used to calculate two multivariate capability measures: the multivariate capability vector Shahriari et al. (1995) and the multivariate capability index MC_{pm} Taam et al. (1993). The proposed software requires no installation, since the user can directly open the .xls file. The spreadsheet interface is easy to use, moreover a set of instructions can be visualized clicking the button “INSTRUCTIONS”. Because a problem with the multivariate measures of capability is the shortage of user-friendly programs we hope that this tool can help practitioners in performing multivariate capability analyses. The software

has been validated using several data set. However, the user should understand that there may be undetected bugs and problems and will be grateful for any feedback with relevant comments and suggestions for improvements. The program mentioned in this article is available from the author writing to: michele.scagliarini@unibo.it.

Acknowledgments

This work has been partly funded by a 2006 MIUR (Italian Ministry for University and Research) grant.

References

- Chen, H. (1994). A multivariate process Capability index over a rectangular solid zone. *Statistica Sinica*, 4, 749-758.
- Hubele, N. F. & Shahriari, H. & Cheng, C.S. (1991). A bivariate process Capability vector. In: J.B. Keats, and D.C. Montgomery (Eds.), *Statistical Process Control in Manufacturing*, Marcel Dekker, New York, 299-310.
- Juran, J. M. & Gryna, F.M. & Bingham, R. S. Jr (1974). *Quality Control Handbook*, McGraw-Hill, New York.
- Kendall, M. (1961) *A course in the Geometry of n-dimensions*. Charles Griffin and Company limited, London.
- Knusel, L. (1998). On the accuracy of statistical distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis*, 26, 375-377.
- Kotz, S. & Lovelace, C. (1998). *Process Capability Indices in Theory and Practice*. Arnold, London.
- McCullough, B. D. & Wilson, B. (2002). On the accuracy of statistical procedures in Microsoft Excel 2000 and Excel XP. *Computational Statistics and Data Analysis*, 40, 713-721.
- Phandnis, M. A. & Elam, M. E. & Fonseca, D.J. & Batson, R.G. & Adams B.M. (2005). A Prototype DSS for PCI Selection, Calculation, and Interpretation. *Proceedings of the 14th Industrial Engineering Research Conference*, <http://www.bama.ua.edu/~melam/pdss.ht>
- Shahriari, H. & Hubele, N. F. & Lawrence, F.P. (1995). A multivariate process capability vector. *Proceedings of the 4th Industrial Engineering Research Conference*, Institute of Industrial Engineers, 304-309.
- Taam, W. & Subbaiah, P. & Liddy, W. J. (1993). A note on multivariate capability indices. *Journal of Applied Statistics*, 20(3) 339-351.
- Wang, F. K, & Hubele, N.F. & Lawrence, F.P. & Miskulin, J.D. & Shahriari, H. (2000). Comparison of three multivariate process capability indices. *Journal of Quality Technology*, 32(3), 263-275.

Translations, Ephemerals, & Biographies

Jacques Salomon Hadamard and the Use of Symbols in Teaching Differential Calculus

Daniel S. Drucker, Claude Schochet
Department of Mathematics

John Cuzzocrea, Shlomo S. Sawilowsky
College of Education
Wayne State University

Scripta Universitatis, edited by Albert Einstein and first published in 1923, played a significant role in the establishment of Hebrew University in Jerusalem. Articles appeared on the left half of the journal in the author's chosen language and they were translated into Hebrew on the right half. The inaugural issue contained an article by the French mathematician Jacques Hadamard (8 December 1865 – 17 October 1963). Y. Wolfson of Kharkov translated it into Hebrew. An English translation is presented here, along with scans of the original first pages that were published in French and Hebrew. Documents pertaining to the origins of *Scripta* were translated from Hebrew into English in *JMASM* (2007, 6(1), p. 350-354).

Key words: mathematics education, differential calculus

Teaching the Concept of Differentials
by J. Hadamard, Paris

Poincaré, speaking at the Pedagogical Museum of Paris in 1904, stated that there was good reason to think in terms of derivatives rather than in terms of differentials. It seems to me that it would be beneficial for the teaching profession to conform resolutely to this principle and to abandon the rather complicated explanations which are classically given via the symbol d . As far as the first differential is concerned, it's OK; I can understand the equation

$$dy = f'(x) dx \quad (1)$$

or

$$dz = p dx + q dy \quad (1')$$

as corresponding to the respective approximate equality

$$\Delta y = f'(x) \Delta x \quad (2)$$

or

$$\Delta z = p \Delta x + q \Delta y \quad (2')$$

in which Δx , Δy , and Δz are infinitesimally small increments. But the second differential! Like everybody, I have read the account of the differential of the independent variable which must be constant (and which is, moreover, necessarily variable because it is infinitely small). If I decided not to cover these concepts in the classes that I have taught in elementary differential calculus, I confess it was because I only half understood them myself.

I know that they must nevertheless be understandable, and if I had needed them, for example, in geometric applications of differential calculus in my research, then I would have mastered them. However, clearly,

Daniel S. Drucker is a Professor of Mathematics. His research interests are in differential geometry, Lie groups, number theory, and applied linear algebra. Email him at drucker@math.wayne.edu. Claude Schochet is a Professor of Mathematics. His research interests are in algebraic topology and C*-algebras. Email him at schochet@math.wayne.edu. John Cuzzocrea is a Lecturer in Research and Statistical Methods. Email: jcuzzocrea@hotmail.com. Shlomo Sawilowsky is an Assistant Dean in the College of Education. His research interests are in rank-based nonparametric statistics. Email him at shlomo@wayne.edu.

such was not the case. I studied infinitesimal geometry like everyone else, having no special difficulty with second derivatives and never giving a thought to leaving the differential of the independent variable constant.

What, then, is the significance of equation (1)? Quite simply, it is that if x and therefore $y = f(x)$ are functions of an arbitrary dependent variable u , then, regardless of the relationship between x and u (provided that dx/du exists), one has

$$\frac{dy}{du} = f'(x) \frac{dx}{du}, \quad (3)$$

which is simply the chain rule.

What does equation (1') mean? Simply stated, if x , y , and hence $z = f(x, y)$, are expressed as functions of an arbitrary dependent variable u , then one has

$$\begin{aligned} \frac{dz}{du} &= \frac{\partial z}{\partial x} \frac{dx}{du} + \frac{\partial z}{\partial y} \frac{dy}{du} \\ &= p \frac{dx}{du} + q \frac{dy}{du}. \end{aligned} \quad (3')$$

Such is the unique meaning of the equations (1) and (1'). Because the equations (3) and (3') hold for any independent variable u in terms of which the other variables are expressed, the variable u can be suppressed. The special advantage of differential notation is that it does not have to specify which variable is to be considered as independent.

What is the meaning of the equation

$$d^2y = f'(x) d^2x + f''(x) dx^2 \quad (4)$$

or

$$\begin{aligned} d^2z &= p d^2x + q d^2y + r dx^2 \\ &+ 2s dx dy + t dy^2? \end{aligned} \quad (4')$$

Simply that

$$\frac{d^2y}{du^2} = f'(x) \frac{d^2x}{du^2} + f''(x) \left(\frac{dx}{du} \right)^2 \quad (5)$$

or

$$\begin{aligned} \frac{d^2z}{du^2} &= p \frac{d^2x}{du^2} + q \frac{d^2y}{du^2} + r \left(\frac{dx}{du} \right)^2 \\ &+ 2s \frac{dx}{du} \frac{dy}{du} + t \left(\frac{dy}{du} \right)^2 \end{aligned} \quad (5')$$

where the variables have been expressed as functions of the parameter u and we still have $y = f(x)$ or $z = f(x, y)$.

Finally, what does the equation

$$d^2z = r dx^2 + 2s dx dy + t dy^2 \quad (6)$$

mean? In my opinion, nothing at all. The first two terms on the right-hand side of (5') disappear when x and y are *linear* functions of u , as happens in the proof of Taylor's theorem, the only place to my knowledge where formula (6) would arise. Aside from this, I don't see what one can get out of (6), except perhaps for one or two false ideas.

References

Hadamard, J. (1923). La notion de différentielle dans l'enseignement. *Scripta Universitatis Atque Bibliothecae Hierosolymitanarum: Mathematica et Physica / Volumen I*, 4. Musag HaDiferentzial B'Horah. Kisvei HaUniversitah V'Bais-HaSofrim B'Yerushalyim: Matimatikah U'F'sika, Crach Aleph, 4.

Appendices

Following are scans of Hadamard's article in the original French and the Hebrew translation, an autographed letter signed (ALS) written by Hadamard to Monsieur E. F. Katz, and scans of title pages of selected texts by Hadamard.

La notion de différentielle dans l'enseignement.

Par

J. Hadamard, Paris.

POINCARÉ, dans sa conférence prononcée au Musée pédagogique de Paris en 1904, déclarait déjà qu'il y avait lieu de penser en dérivées et non en différentielles. Il me semble utile pour l'enseignement de se conformer résolument à ce principe et d'abandonner les explications assez compliquées qui sont classiquement données sur le symbole d . Pour la différentielle première, passe encore: Je puis comprendre l'égalité

$$(1) \quad dy = f'(x) dx$$

ou

$$(1') \quad dz = p dx + q dy$$

en relation avec l'égalité approchée

$$(2) \quad \Delta y = f'(x) \Delta x$$

ou

$$(2') \quad \Delta z = p \Delta x + q \Delta y$$

dans laquelle Δx , Δy , Δz sont des accroissements infiniment petits. Mais la différentielle seconde! J'ai lu, comme tout le monde, l'histoire de la différentielle de la variable indépendante qui doit être constante (et qui est d'ailleurs forcément variable puisque infiniment petite). Si je me suis décidé à ne pas exposer ces considérations dans les cours que j'ai eu l'occasion de professer sur les débuts du Calcul différentiel, c'est que j'avoue ne les avoir qu'à moitié comprises moi-même.

From the *JMASM* Archives.

מושג הדיפרנציאל בהוראה.

י. הדמר, פריס.

תרגם י. ולפסון, חרקוב.

פואנקרה בשעתו חזה דעתו, בדרשה שדרש במחזאון הפדגוגי של פריס בשנת 1904, כי המחשבה המדעית צריכה להשתמש בתורידיות ולא בדיפרנציאלים. ולדעתי יפה הוא זה להוראה, אם נסתגל בהחלט לפרינציפיון הזה ונסתלק מהבאורים הקלסיים המרכבים למדי שנותנים לסימבול d . אשר לדיפרנציאל הראשון, עדיין, בכי טוב הוא. וכולני להבין את השויון

$$dy = f'(x) dx \quad (1)$$

או

$$dz = p dx + q dy \quad (1')$$

בקשור עם השויון המקורב

$$\Delta y = f'(x) \Delta x, \quad (2)$$

או

$$\Delta z = p \Delta x + q \Delta y \quad (2')$$

שבו Δz , Δy , Δx הם הוספות קטנות בלי-סוף. אבל הדיפרנציאל השני! אף אני קראתי, נבל העולם כולו, את המעשיה על הדיפרנציאל של המשתנה החפשי שחייב הוא להיות קבוע ועומד דרך אנב משתנה ביותר, מפני שהוא קטן עד לאין-סוף). ואשר גמרתי

M 4°

From the JMASM Archives.

From the *JMASM* Archives.

Paris, 5 Novembre 1932
Bne Jean Polent (14^{arr^e})

Monsieur,
C'est avec grand plaisir que j'ai écrit pour dire —
à vous, j'en suis sûr,

Ne wieder Krieg!

Puisse ce cri être porté et entendu par le plus d'hommes
possible, dans votre pays et dans le mien!

Croyez à mes sentiments très dévoués

J. Hadamard

Membre de l'Académie des Sciences de Paris
Professeur au Collège de France

JACQUES SALOMON HADAMARD

COLUMBIA UNIVERSITY IN THE CITY OF NEW YORK
PUBLICATION NUMBER FIVE
OF THE ERNEST KEMPTON ADAMS FUND FOR PHYSICAL RESEARCH
ESTABLISHED DECEMBER 17TH, 1904

FOUR LECTURES ON MATHEMATICS

DELIVERED AT COLUMBIA UNIVERSITY
IN 1911

BY

J. HADAMARD

MEMBER OF THE INSTITUTE, PROFESSOR IN THE COLLÈGE DE FRANCE AND IN THE ÉCOLE POLYTECHNIQUE,
LECTURER IN MATHEMATICS AND MATHEMATICAL PHYSICS IN COLUMBIA UNIVERSITY FOR 1911

NEW YORK
COLUMBIA UNIVERSITY PRESS
1915

From the JMASM Archives.

ZEROS DES FONCTIONS HOLOMORPHES.

THEOREME.- Tout zéro d'une fonction holomorphe (dans une aire où elle n'est pas identiquement nulle) :

1°/ a un ordre de multiplicité fini et déterminé.

2°/ est isolé, c'est-à-dire à une distance finie de tout autre zéro.

1°/ Soit $f(z) = A_0 + A_1(z-a) + \dots + A_n(z-a)^n + \dots$

le développement de la fonction autour du point ordinaire a . Si ce point est un zéro, A_0 est nul. Il peut en être de même de quelques-uns des coefficients suivants, mais non de tous (puisque f n'est pas identiquement nul). Si A_n est le premier coefficient non nul, on a

$$\begin{aligned} f(z) &= (z-a)^n \varphi(z) \\ \varphi(a) &= A_n \neq 0. \end{aligned}$$

L'ordre de multiplicité est alors n .

2°/ La série qui représente $\varphi(z)$ est une série de puissances qui admet même cercle de convergence que $f(z)$, c'est donc une fonction holomorphe dans le même cercle, qui est différente de zéro pour $z = a$. Cette fonction étant continue, on trouvera un nombre ρ positif tel que la fonction ne s'annule pas dans le cercle de rayon ρ et de centre a ; $\varphi(z)$ n'admet pas de zéro à l'intérieur de ce cercle, et $\varphi(z)$ n'admet pas à l'intérieur de ce cercle d'autre zéro que $z = a$.

Exercice.- Trouver une limite inférieure de ρ , connaissant $\varphi(z) = A_n + A_{n+1}(z-a) + A_{n+p}(z-a)^p + \dots$ ($A_n \neq 0$)

(question analogue à la recherche d'une limite inférieure des racines positives d'une équation).

On écrira $\varphi(z) = A_n - \varphi_1(z)$

avec pour $\varphi_1(z)$ une majorante de la forme

$$\varphi_1(z) << M \left(\frac{1}{1 - \frac{z}{R}} - 1 \right) = \frac{Mz}{R-z}$$

193.- On déduit de là le

Théorème.- Une fonction holomorphe dans une aire limitée en tous sens et sur le contour de cette aire (1) ne peut y admettre qu'un nombre fini de zéros.

Car les zéros, supposés en nombre infini, admettraient (1^{re} année n° 56) soit dans l'aire, soit sur son contour, un point d'accumulation, ce que nous venons de démontrer être impossible.

Remarque.- Mais une fonction holomorphe dans tout le plan (fonction entière) y admet, du moins en général (exemple $\sin z$, exemple contraire e^z), une infinité de zéros.

(1)- Cette partie de l'hypothèse est indispensable. Exemple : $\sin \frac{1}{z}$, dans une demi circonférence ayant son centre à l'origine et limitée à l'axe imaginaire (fig. ci-contre), a une infinité de zéros $x = \frac{1}{k\pi}$. Le point singulier est l'origine, qui est un point essentiel (voir plus loin n° 196).

AN ESSAY ON
The Psychology of Invention
in the
Mathematical Field

BY
JACQUES HADAMARD

PRINCETON
PRINCETON UNIVERSITY PRESS
1945

From the *JMASM* Archives.

It's Back!

Design and Analysis of Time-Series Experiments

(with a new Introduction by the first author)

Gene V Glass, Arizona State University

Victor L. Willson, Texas A&M University

John M. Gottman, The Gottman Institute, Seattle, Washington

Hailed as a landmark in the development of experimental methods when it appeared in 1975, *Design and Analysis of Time-Series Experiments* is available again after several years of being out of print.

Gene V Glass, Victor L. Willson and John M. Gottman have carried forward the design and analysis of perhaps the most powerful and useful quasi-experimental design identified by their mentors in the classic Campbell & Stanley text *Experimental and Quasi-experimental Design for Research* (1966). In an era when governments seek to resolve questions of experimental validity by fiat and the label "Scientifically Based Research" is appropriated for only certain privileged experimental designs, nothing could be more appropriate than to bring back the classic text that challenges doctrinaire opinions of proper causal analysis.

Glass, Willson & Gottman introduce and illustrate an armamentarium of interrupted time-series experimental designs that offer some of the most powerful tools for discovering and validating causal relationships in social and education policy analysis. Drawing on the ground-breaking statistical analytic tools of Box & Jenkins, the authors extend the comprehensive autoregressive-integrated-moving-averages (ARIMA) model to accommodate significance testing and estimation of the effects of interventions into real world time-series. Designs and full statistical analyses are richly illustrated with actual examples from education, behavioral psychology, and sociology.

"...this book will come to be viewed as a true landmark. ... [It] should stand the test of time exceedingly well." ~ James A. Walsh (*Educational & Psychological Measurement*, 1975)

"Ordinary least squares estimation is usually inapplicable because of autoregressive error.... Glass, Willson, and Gottman have assembled the best approach." ~Donald T. Campbell

Publication Date:

Winter 2009

ISBN's:

Paperback: 978-1-59311-980-5

Hardcover: N/A

Price:

Paperback: \$39.99

Hardcover:

Trim Size: 6 X 9

Subject:

Education, Statistics

Special Price: \$25.99 paperbacks plus s/h

Book URL: <http://www.infoagepub.com/products/content/p489c9049a428d.php>

It's Back in Paperback!

Statistical Theories of Mental Test Scores

by **Frederic M. Lord** and **Melvin R. Novick**

A classic returns.

...pioneering work...
...comprehensive...
...classic...
...definitive...
...unquestioned status and authority...

Tatsuoka was right:

"This comprehensive and authoritative work is a major contribution to the literature of test theory. Without doubt it is destined to become a classic in the field." ~ Maurice Tatsuoka (1971)

One of the most important books in the history of psychometrics has been virtually unavailable to scholars and students for decades. A gap in the archives of modern test theory is now being filled by the release in paperback for the first time of the classic text, *Statistical Theories of Mental Test Scores*, by the late and honored statisticians and psychometricians, Frederic M. Lord and Melvin R. Novick. No single book since 1968 when Lord & Novick first appeared has had a comparable impact on the practice of testing and assessment.

Information Age Publishing is proud to make this classic text available to a new generation of scholars and researchers.

Publication Date:

Spring 2008

ISBN's:

Paperback: 978-1-59311-934-8

Price:

Paperback: \$59.99

Trim Size: 6 X 9

Subject:

Education, Statistics

<http://www.infoagepub.com/products/content/p4810c9a0891af.php>

Advances in Latent Variable Mixture Models

Edited by **Gregory R. Hancock**, *University of Maryland, College Park*,
and **Karen M. Samuelsen**, *University of Georgia*

The current volume, *Advances in Latent Variable Mixture Models*, contains chapters by all of the speakers who participated in the 2006 CILVR conference, providing not just a snapshot of the event, but more importantly chronicling the state of the art in latent variable mixture model research. The volume starts with an overview chapter by the CILVR conference keynote speaker, Bengt Muthén, offering a “lay of the land” for latent variable mixture models before the volume moves to more specific constellations of topics. Part I, *Multilevel and Longitudinal Systems*, deals with mixtures for data that are hierarchical in nature either due to the data’s sampling structure or to the repetition of measures (of varied types) over time. Part II, *Models for Assessment and Diagnosis*, addresses scenarios for making judgments about individuals’ state of knowledge or development, and about the instruments used for making such judgments. Finally, Part III, *Challenges in Model Evaluation*, focuses on some of the methodological issues associated with the selection of models most accurately representing the processes and populations under investigation. It should be stated that this volume is not intended to be a first exposure to latent variable methods. Readers lacking such foundational knowledge are encouraged to consult primary and/or secondary didactic resources in order to get the most from the chapters in this volume. Once armed with that basic understanding of latent variable methods, we believe readers will find this volume incredibly exciting.

CONTENTS: Editors’ Introduction, *Gregory R. Hancock and Karen M. Samuelsen*. Acknowledgments. Latent Variable Hybrids: Overview of Old and New Models, *Bengt Muthén*. **PART I: Multilevel and Longitudinal Systems.** Multilevel Mixture Models, *Tihomir Asparouhov and Bengt Muthén*. Longitudinal Modeling of Population Heterogeneity: Methodological Challenges to the Analysis of Empirically Derived Criminal Trajectory Profiles, *Frauke Kreuter and Bengt Muthén*. Examining Contingent Discrete Change Over Time with Associative Latent Transition Analysis, *Brian P. Flaherty*. Modeling Measurement Error in Event Occurrence for Single, Non-Recurring Events in Discrete-Time Survival Analysis, *Katherine E. Masyn*. **PART II: Models for Assessment and Diagnosis.** Evidentiary Foundations of Mixture Item Response Theory Models, *Robert J. Mislevy, Roy Levy, Marc Kroopnick, and Daisy Rutstein*. Examining Differential Item Functioning from a Latent Mixture Perspective, *Karen M. Samuelsen*. Mixture Models in a Developmental Context, *Karen Draney, Mark Wilson, Judith Glück, and Christiane Spiel*. Applications of Stochastic Analyses for Collaborative Learning and Cognitive Assessment, *Amy Soller and Ron Stevens*. The Mixture General Diagnostic Model, *Matthias von Davier*. **PART III: Challenges in Model Evaluation.** Categories or Continua? The Correspondence Between Mixture Models and Factor Models, *Eric Loken and Peter Molenaar*. Applications and Extensions of the Two-Point Mixture Index of Model Fit, *C. Mitchell Dayton*. Identifying the Correct Number of Classes in Growth Mixture Models, *Davood Tofghi and Craig K. Enders*. Choosing a “Correct” Factor Mixture Model: Power, Limitations, and Graphical Data Exploration, *Gitta H. Lubke and Jeffrey R. Spies*. About the Contributors.

Publication Date:
Fall 2007

ISBN's:
Paperback: 978-1-59311-847-1
Hardcover: 978-1-59311-848-8

Price:
Paperback: \$39.99
Hardcover: \$73.99

Trim Size: 6 X 9

Subject:
Education

Books of Related Interest:

Structural Equation Modeling: A Second Course

<http://www.infoagepub.com/products/content/1-59311-015-4.php>

2006

Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.95

Real Data Analysis

<http://infoagepub.com/products/content/978-1-59311-565-4.php>

2007

Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.95

Structural Equation Modeling: A Second Course

Edited by **Gregory R. Hancock**, *University of Maryland*
and **Ralph O. Mueller**, *The George Washington University*

A volume in **Quantitative Methods in Education and the Behavioral Sciences:
Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)

"I believe that this volume represents a vital contribution to the field of SEM beyond the introductory level."

From the Preface by
Richard G. Lomax, *The University of Alabama*

This volume is intended to serve as a didactically-oriented resource covering a broad range of advanced topics often not discussed in introductory courses on structural equation modeling (SEM). Such topics are important in furthering the understanding of foundations and assumptions underlying SEM as well as in exploring SEM as a potential tool to address new types of research questions that might not have arisen during a first course. Chapters focus on the clear explanation and application of topics, rather than on analytical derivations, and contain syntax and partial output files from popular SEM software.

CONTENTS: Introduction to Series, *Ronald C. Serlin*. Preface, *Richard G. Lomax*. Dedication. Acknowledgements. Introduction, *Gregory R. Hancock & Ralph O. Mueller*. **Part I: Foundations.** The Problem of Equivalent Structural Models, *Scott L. Hersherberger*. Formative Measurement and Feedback Loops, *Rex B. Kline*. Power Analysis in Covariance Structure Modeling, *Gregory R. Hancock*. **Part II: Extensions.** Evaluating Between-Group Differences in Latent Variable Means, *Marilyn S. Thompson & Samuel B. Green*. Using Latent Growth Models to Evaluate Longitudinal Change, *Gregory R. Hancock & Frank R. Lawrence*. Mean and Covariance Structure Mixture Models, *Phill Gagné*. Structural Equation Models of Latent Interaction and Quadratic Effects, *Herbert W. Marsh, Zhonglin Wen, & Kit-Tai Hau*. **Part III: Assumptions.** Nonnormal and Categorical Data in Structural Equation Modeling, *Sara J. Finney & Christine DiStefano*. Analyzing Structural Equation Models with Missing Data, *Craig K. Enders*. Using Multilevel Structural Equation Modeling Techniques with Complex Sample Data, *Laura M. Stapleton*. The Use of Monte Carlo Studies in Structural Equation Modeling Research, *Deborah L. Bandalos*. About the Authors.

Also Available:

Multilevel Modeling of Educational Data

2008 Paperback ISBN: 978-1-59311-684-2 \$39.99 Hardcover ISBN: 978-1-59311-685-9 \$73.99

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.99

Publication Date:
2005

ISBN's:
Paperback: 1-59311-014-6
Hardcover: 1-59311-015-4

Price:
Paperback: \$39.99
Hardcover: \$73.99

Subject:
Education, Statistics

Series URL: <http://www.infoagepub.com/products/series/serlin.html>

Multilevel Modeling of Educational Data

Edited by **Ann A. C'Connell**, *Ohio State University*
and **D. Betsy McCoach**, *University of Connecticut*

A volume in **Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching**

Series Editor **Ron Serlin**, *University of Wisconsin*

(sponsored by the Educational Statisticians, SIG)

Multilevel Modeling of Educational Data, co-edited by Ann A. O'Connell, Ed.D., and D. Betsy McCoach, Ph.D., is the next volume in the series: *Quantitative Methods in Education and the Behavioral Sciences: Issues, Research and Teaching* (Information Age Publishing), sponsored by the Educational Statisticians' Special Interest Group (Ed-Stat SIG) of the American Educational Research Association. The use of multilevel analyses to examine effects of groups or contexts on individual outcomes has burgeoned over the past few decades. Multilevel modeling techniques allow educational researchers to more appropriately model data that occur within multiple hierarchies (i.e.- the classroom, the school, and/or the district). Examples of multilevel research problems involving schools include establishing trajectories of academic achievement for children within diverse classrooms or schools or studying school-level characteristics on the incidence of bullying. Multilevel models provide an improvement over traditional single-level approaches to working with clustered or hierarchical data; however, multilevel data present complex and interesting methodological challenges for the applied education research community.

In keeping with the pedagogical focus for this book series, the papers this volume emphasize applications of multilevel models using educational data, with chapter topics ranging from basic to advanced. This book represents a comprehensive and instructional resource text on multilevel modeling for quantitative researchers who plan to use multilevel techniques in their work, as well as for professors and students of quantitative methods courses focusing on multilevel analysis. Through the contributions of experienced researchers and teachers of multilevel modeling, this volume provides an accessible and practical treatment of methods appropriate for use in a first and/or second course in multilevel analysis. A supporting website links chapter examples to actual data, creating an opportunity for readers to reinforce their knowledge through hands-on data analysis. This book serves as a guide for designing multilevel studies and applying multilevel modeling techniques in educational and behavioral research, thus contributing to a better understanding of and solution for the challenges posed by multilevel systems and data.

CONTENTS: **Series Introduction**, *Ronald C. Serlin*. **Acknowledgements**. Part I: **Design Contexts for Multilevel Models**. Introduction, *Ann A. O'Connell and D. Betsy McCoach*. The Use of National Datasets for Teaching and Research, *Laura M. Stapleton and Scott L. Thomas*. Using Multilevel Modeling to Investigate School Effects, *Xin Ma, Lingling Ma, and Kelly D. Bradley*. Modeling Growth Using Multilevel and Alternative Approaches, *Janet K. Holt*. Cross-Classified Random Effects Models, *S. Natasha Beretvas*. Multilevel Logistic Models for Dichotomous and Ordinal Data, *Ann A. O'Connell, Jessica Goldstein, H. Jane Rogers, and C. Y. Joanne Peng*. Part II: **Planning and Evaluating Multilevel Models**. Evaluation of Model Fit and Adequacy, *D. Betsy McCoach and Anne C. Black*. Power, Sample Size, and Design, *Jessica Spybrook*. Part III: **Extending the Multilevel Framework**. Multilevel Methods for Meta-Analysis, *Sema A. Kalaian and Rafa M. Kasim*. Multilevel Measurement Modeling, *Kihito Kamata, Daniel J. Bauer, and Yasuo Miyazaki*. Part IV: **Mastering the Technique**. Reporting Results from Multilevel Analyses, *John M. Ferron, Kristin Y. Hogarty, Robert F. Dedrick, Melinda R. Hess, John D. Niles, and Jeffrey D. Kromrey*. Software Options for Multilevel Models, *J. Kyle Roberts and Patrick McLeod*. Estimation Procedures for Hierarchical Linear Models, *Hariharan Swaminathan and H. Jane Rogers*.

Also Available:

Real Data Analysis

2007 Paperback ISBN: 978-1-59311-564-7 \$39.99 Hardcover ISBN: 978-1-59311-565-4 \$73.95

Structural Equation Modeling: A Second Course

2005 Paperback ISBN: 1-59311-014-6 \$39.99 Hardcover ISBN: 1-59311-015-4 \$73.95

Publication Date:
Spring 2008

ISBN's:
Paperback: 978-1-59311-684-2
Hardcover: 978-1-59311-685-9

Price:
Paperback: \$39.99
Hardcover: \$73.99

Subject:
Education, Statistics

Series URL: <http://www.infoagepub.com/products/series/serlin.html>

Book URL: <http://www.infoagepub.com/products/content/p478cb9504908a.php>

IAP - Information Age Publishing, PO Box 79049, Charlotte, NC 28271
tel: 704-752-9125 fax: 704-752-9113 URL: www.infoagepub.com

Instructions For Authors

Follow these guidelines when submitting a manuscript:

1. *JMASM* uses a modified American Psychological Association style guideline.
2. Submissions are accepted via e-mail only. Send them to the Editorial Assistant at ea@edstat.coe.wayne.edu. Provide name, affiliation, address, e-mail address, and 30 word biographical statements for all authors in the body of the email message.
3. There should be no material identifying authorship except on the title page. A statement should be included in the body of the e-mail that, where applicable, indicating proper human subjects protocols were followed, including informed consent. A statement should be included in the body of the e-mail indicating the manuscript is not under consideration at another journal.
4. Provide the manuscript as an external e-mail attachment in MS Word for the PC format only. (Wordperfect and .rtf formats may be acceptable - please inquire.) Please note that Text (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.
5. The text maximum is 20 pages double spaced, not including tables, figures, graphs, and references. Use 11 point Times Roman font.
6. 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.
7. The manuscript should contain an Abstract with a 50 word maximum, following by 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.
8. 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.
9. 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.
10. *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 spacebar twice after a period.

Print Subscriptions

Print subscriptions including postage for professionals are US \$95 per year; for graduate students are US \$47.50 per year; and for libraries, universities, and corporations are US \$195 per year. Subscribers outside of the US and Canada pay a US \$10 surcharge for a additional postage. Online access is currently free at <http://tbf.coe.wayne.edu/jmasm>. Mail subscription requests with remittances to JMASM, P. O. Box 48023, Oak Park, MI, 48237. Email journal correspondence, other than manuscript submissions, to jmasm@edstat.coe.wayne.edu.

Notice To Advertisers

Send requests for advertising information to jmasm@edstat.coe.wayne.edu.

The easy way to find open access journals

www.doaj.org

The Directory of Open Access Journals covers free, full text, quality controlled scientific and scholarly journals. It aims to cover all subjects and languages.

Aims

- Increase visibility of open access journals
- Simplify use
- Promote increased usage leading to higher impact

Scope

The Directory aims to be comprehensive and cover all open access scientific and scholarly journals that use a quality control system to guarantee the content. All subject areas and languages will be covered.

In DOAJ browse by subject

Agriculture and Food Sciences
Biology and Life Sciences
Chemistry
General Works
History and Archaeology
Law and Political Science
Philosophy and Religion
Social Sciences

Arts and Architecture
Business and Economics
Earth and Environmental Sciences
Health Sciences
Languages and Literatures
Mathematics and statistics
Physics and Astronomy
Technology and Engineering

Contact

Lotte Jørgensen, Project Coordinator
Lund University Libraries, Head Office
E-mail: lotte.jorgensen@lub.lu.se
Tel: +46 46 222 34 31

Funded by

www.soros.org

Hosted by

LUND
UNIVERSITY
www.lu.se