DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 1

5-1-2017

Robust ANCOVA: Confidence Intervals That Have Some Specified Simultaneous Probability Coverage When There Is Curvature And Two Covariates

Rand Wilcox
University of Southern California, rwilcox@usc.edu

Recommended Citation

Wilcox, R. (2017). Robust ANCOVA: Confidence intervals that have some specified simultaneous probability coverage when there is curvature and two covariates. Journal of Modern Applied Statistical Methods, 16(1), 3-19. doi: 10.22237/jmasm/1493596800

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 3-19.
doi: 10.22237/jmasm/1493596800

ISSN 1538 − 9472

Robust ANCOVA: Confidence Intervals That Have Some Specified Simultaneous Probability Coverage When There Is Curvature And Two Covariates

Rand Wilcox
University of Southern California
Los Angeles, California

Consider the commonly occurring situation where the goal is to compare two independent groups and there are two covariates. Let $M_j(X)$ be some conditional measure of location for the j^{th} group associated with some random variable Y given $X = (X_1, X_2)$. The goal is to H_0: $M_1(X) = M_2(X)$ for each $X \in \Omega$ in a manner that controls the probability of one or more Type I errors. An extant technique (method M_1 here) addresses this goal without making any parametric assumption about $M_j(X)$. However, a practical concern is that it does not provide enough detail regarding where the regression surfaces differ, due to using a very small number of covariate points, which can result in relatively low power. Method M_2 was proposed for testing the global hypothesis H_0: $M_1(X) = M_2(X)$ for all $X \in \Omega$, which offers a distinct power advantage over method M_1. It uses the deepest half of the covariate points rather than small number of points used by method M_1. However, method M_2 does not provide any details about which covariate points yield a significant result. A multiple comparison procedure is proposed that deals with this shortcoming of method M_2, and simultaneously it can provide higher power than method M_1.

Keywords: ANCOVA, trimmed mean, smoothers, Well Elderly 2 study

Introduction

Consider the common situation where the goal is to compare two independent groups based on two covariates. The classic ANCOVA (analysis of covariance) method assumes that

$$Y_j = \beta_{0j} + \beta_1 X_{1j} + \beta_2 X_{2j} + \varepsilon \,, \tag{1}$$

Rand Wilcox is an Professor in the Department of Psychology. Email him at: rwilcox@usc.edu.

where β_{0j}, β_1 and β_2 are unknown parameters estimated via least squares regression and ε is a random variable having a normal distribution with mean zero and unknown variance σ^2. So the regression planes are assumed to be parallel and the groups can be compared by testing

$$\mathrm{H}_0 : \beta_{01} = \beta_{02}, \tag{2}$$

the hypothesis that the intercepts are equal. It is well known, however, that least squares regression is not robust (e.g., Staudte and Sheather, 1990; Maronna et al., 2006; Heritier et al., 2007; Hampel et al., 1986; Huber and Ronchetti, 2009; Wilcox, 2012). A practical consequence is that power can be relatively low even under a small departure from normality. Moreover, even a single outlier can yield a poor fit to the bulk of the points when using least squares regression.

Another concern with the classic ANCOVA model is that two types of homoscedasticity are assumed. The first is that for each group, the variance of the error term does not depend on the value of the covariate. If this assumption is violated the wrong standard error is being used (e.g., Long & Ervin, 2000). A seemingly natural way of justifying a homoscedastic error term is to test the assumption that it is indeed homoscedastic. However, Ng and Wilcox (2011) found that this strategy is unsatisfactory. The problem is that methods for testing the homoscedasticity assumption do not have enough power to detect situations where heteroscedasticity is a practical concern. The second homoscedasticity assumption is that the variance of the error term is the same for both groups. Violating these assumptions can result in poor control over the Type I error probability.

Yet another fundamental concern with (1) is that the true regression surfaces are assumed to be planes. Presumably this is a reasonable approximation in some situations, but experience with smoothers (e.g., Hastie & Tibsherani, 1990; Wilcox, 2012) made it clear that often this is not the case. When there is curvature, using some obvious parametric regression model might suffice. (For example, include a quadratic term.) It is known that this approach can be inadequate, which has led to a substantial collection of nonparametric regression methods, often called smoothers, for dealing with curvature in a more flexible manner (e.g., Härdle, 1990; Efromovich, 1999; Eubank, 1999; Fox, 2001; Györfi, et al., 2002).

One more limitation of the classic model is the assumption that the regression surfaces are parallel. The assumption that the slope parameters are equal could be tested, but it is unclear when such a test has enough power to detect situations where this assumption is violated to the point that it makes a practical difference.

Let $M_j(X)$ be some conditional measure of location associated with Y given $X = (X_1, X_2)$, where $M_j(X)$ is some unknown function. Here, the model given by (1) is replaced with the less restrictive model

$$Y_j = M_j\left(X\right) + \lambda\left(X\right)\varepsilon_j, \tag{3}$$

where $\lambda(X)$ is some unknown function used to model heteroscedasticity. The random variable ε_j has some unknown distribution with variance σ_j^2. So unlike the classic approach where it is assumed that

$$M_j\left(X\right) = \beta_{0j} + \beta_{1j}X_1 + \beta_{2j}X_2,$$

no parametric model for $M_j(X)$ is specified and $\sigma_1 = \sigma_2$ is not assumed. In particular it is not assumed that the regression surfaces are parallel.

Let $X_1, \ldots, X_K$ be K covariate points that are chosen empirically in a manner to be described. The goal here is to test the K hypotheses

$$\mathrm{H}_0 : M_1\left(X_k\right) = M_2\left(X_k\right) \tag{4}$$

for each $k = 1, \ldots, K$ such that the probability of one or more Type I errors is approximately equal to α. The focus is on situations where $M_j(X)$ is a trimmed mean, but the basic strategy underlying the proposed approach (method M_3 in the so-named section) can in principle be extended to other robust measures of location.

Wilcox (2012) suggested a simple method for testing (4) for each $k = 1, \ldots, K$ when the covariate points are chosen based on how deeply they are nested within the cloud of covariate points (this is method M_1 in the so-named section). The K points are chosen to include the point in the first group having the deepest half space depth plus the points on the .5 depth contour. This typically results in using a fairly small number of covariate points where the corresponding Y values are compared based on a robust measure of location. Among the K tests that are performed, the probability of one or more Type I errors can be controlled using some improvement on the Bonferroni method (e.g., Hommel, 1988; Hochberg, 1988). However, it is not clear when this relatively simple approach will choose covariate values that are likely to detect true differences between the groups. Another concern is that important details about where the groups differ will be missed due to using a small number of covariate points.

A way of dealing with this issue is to select a larger collection of covariate points. The strategy here is to use the deepest half of the covariate points in the first group. But as K increases, an obvious concern is the negative impact this will have on power when using the methods derived by Hommel (1988) and Hochberg (1988). A method that controls the false discovery rate when dealing with dependent test statistics (e.g., Benjamini &Yekutieli, 2001) suffers from the same concern. Wilcox (2016) derived a method for testing the global hypothesis that (4) is true for the deepest half of the covariate points in the first group (this is method M_2 in the so-named section). However, when this method rejects, it provides virtually no information about which of the individual hypotheses can be rejected.

The goal here is to suggest a method for controlling the probability of one or more Type I errors when testing the K hypotheses given by (4). Like method M_2, the deepest half of the covariate points is used. But rather than use the methods derived by Hommel (1988) and Hochberg (1988), an alternative technique is suggested that has a certain similarity to using a Studentized maximum modulus distribution.

Description of the Methods

Let (Y_{ij},X_{ij}) (i = 1, ..., n_j; j = 1, 2) be a random sample from the j^{th} group. The methods compared here are based in part on a method derived by Yuen (1974) for comparing the population trimmed means of two independent groups. To describe it, momentarily ignore the covariates and consider the goal of testing

$$\mathrm{H}_0 : \mu_{t1} = \mu_{t2}, \tag{5}$$

the hypothesis that two independent groups have equal population trimmed means. For the j^{th} group (j = 1, 2), let $Y_{(1)j} \leq \ldots \leq Y_{(n_j)j}$ denote the Y_{ij} values written in ascending order. For some $0 \leq \gamma < .5$, the γ-trimmed mean for the j^{th} group is

$$\bar{Y}_j = \frac{1}{n_j - 2g_j}\left(Y_{(g_j+1)j} + \cdots + Y_{(n_j - g_j)j}\right)$$

where $g_j = [\gamma n_j]$ is the greatest integer less than or equal to γn_j. Here the focus is on $\gamma = .2$, a 20% trimmed mean. Under normality, this choice has good efficiency relative to the sample mean (Rosenberger & Gakso, 1983). Moreover, the sample 20% trimmed mean enjoys certain theoretical advantages. First, it has a reasonably high breakdown point, which refers to the proportion of values that must be altered to destroy it. Asymptotic results and simulations indicate that it reduces substantially concerns about the impact of skewed distributions on the probability of a Type I error (e.g., Wilcox, 2012). This is not to suggest that 20% trimming is always the optimal choice: clearly this is not the case. The only suggestion is that it is a reasonable choice among the many robust estimators that might be used.

Winsorizing the Y_{ij} values refers to setting

$$\begin{aligned}W_{ij} &= Y_{(g_j+1)}, \text{ if } Y_{ij} \le Y_{(g_j+1)}\\ W_{ij} &= Y_{ij}, \text{ if } Y_{(g_j+1)} < Y_{ij} < Y_{(n_j-g_j)}\\ W_{ij} &= Y_{(n_j-g_j)}, \text{ if } Y_{ij} \ge Y_{(n_j-g_j)}\end{aligned}$$

The Winsorized sample mean corresponding to group j is the mean based on the Winsorized values, and the Winsorized variance, s_{wj}^2, is the usual sample variance, again based on the Winsorized values.

Let $h_j = n_j - 2g_j$. That is, h_j is the number of observations left in the j^{th} group after trimming. Let

$$d_j = \frac{(n_j - 1)s_{wj}^2}{h_j(h_j - 1)}.$$

Yuen's test statistic is

$$T_y = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{d_1 + d_2}}$$

The null distribution is taken to be a Student's t distribution with degrees of freedom

$$\hat{\upsilon} = \frac{(d_1 + d_2)^2}{C},$$

where

$$C = \frac{d_1^2}{h_1} + \frac{d_2^2}{h_2}$$

Method M_1

Method M_1 was described in Wilcox (2012, section 11.11.3). A complete description of the many computational details is not provided here, but an outline of the method is provided with the goal of explaining how it differs from methods M_2 and M_3.

Momentarily consider a single covariate point, X. For fixed j, method M_1 estimates $M_j(X)$ using the Y_{ij} associated with the X_{ij} points that are close to X. More precisely, for the jth group, compute a robust covariance matrix based on $X_{ij}(i = 1, \ldots, n_j)$. There are many ways of computing a robust covariance matrix with no single estimator dominating. Here a skipped covariance matrix is used, which is computed as follows. For fixed j, outliers among the X_{ij} points are identified using a projection-type multivariate outlier detection technique (e.g., Wilcox, 2012, section 6.4.9). These outliers are removed and the usual covariance matrix is computed using the remaining data.

Next, compute robust Mahalanobis distances for each covariate point based on the robust covariance matrix just described, with X taken to be the center of the data. The point X_{ij} is said to be close to X if its robust Mahalanobis distance is small, say less than or equal to f, which is called the span. Generally, $f = .8$ performs reasonably well when the goal is to approximate the regression surface. Of course exceptions are encountered, but henceforth $f = .8$ is assumed. Let $P_j(X)$ be the subset of $\{1, 2, ..., n_j\}$ that indexes the X_{ij} values such that the Mahalanobis distance associated with X_{ij} is less than or equal to f. Let $N_j(X)$ be the cardinality of the set $P_j(X)$ and let $M_j(X)$ denote the 20% trimmed mean based on the Y_{ij} values for which $i \in P_j(X)$. Then for the single point X, (4) can be tested by applying Yuen's method with the Y_{ij} values for which $i \in P_j(X)$ provided both $N_1(X)$ and $N_2(X)$ are not too small. Following Wilcox (2012), this is taken to mean that Yuen's method can be applied if simultaneously $N_1(X) \geq 12$ and $N_2(X) \geq 12$, in which case the two groups are said to be comparable at X.

Consider the issue of choosing covariate points where the regression surfaces will be compared. For the first group, compute how deeply each X_{i1} is nested within the cloud of covariate points ($i = 1, \ldots, n_j$). This is done with a projection-type method that is similar to an approach discussed by Donoho and

Gasko (1992). The many computational details are not described and are not particularly important for present purposes. Here it is merely noted that an approximation of halfspace depth is used, which is described in Wilcox (2012, section 6.2.3) and labeled approximation A_1. Consider the deepest point as well as those on the polygon containing the central half of the data. (Liu et al., 1999, call this polygon the .5 depth contour.) Method M_1 applies Yuen's method at each of these points provided the regression surfaces are comparable at these points as previously defined. The probability of one or more Type I errors is controlled using the method in Hochberg (1988).

Method M_2

There are several positive features of method M_1 but some negative features as well. First, Yuen's method for comparing trimmed means has been studied extensively and appears to perform relatively well in terms of both Type I errors and power. The method for choosing the covariate values seems reasonable in the sense that it uses points that are nested deeply within the cloud of covariate points, which reflect situations where the regression surfaces are comparable. Roughly, deeply nested points correspond to situations where the regression surfaces can be estimated in a relatively accurate manner. If a point X is not deeply nested in the cloud of covariate values, finding a sufficiently large number of other points that are close to X might be impossible.

But a concern with M_1 is that perhaps true differences might be missed because typically a relatively small number of covariate points are used. In the Illustration to follow, only three covariate points are used by M_1, with sample sizes 187 and 228. Method M_2 deals with this concern in the following manner. First, it computes the projection depth for each X_{i1} (the i^{th} covariate vector in group 1) in the same manner as method M_1. Let the set $\{X_1, \ldots, X_K\}$ indicate the deepest half of the points in the first group. Points where the regression surfaces are not comparable (i.e., $N_1(X) < 12$ or $N_2(X) < 12$) are discarded. Because K can be relatively large, it is approximately equal to $n_1/2$, controlling the probability of one or more Type I errors via Hochberg's method or Hommel's method is likely to have relatively low power.

The reason for choosing the deepest half of the covariate points, rather than some larger proportion, is that typically the regression surfaces are comparable at all K points when the sample sizes for both groups are greater than or equal to 50. For a larger proportion of points, this is often not the case. There are, of course, many other variations. Some other measure of depth might be used or one could use all of the covariate points where the regression surfaces are comparable.

Method M_2 proceeds in the same manner as method M_1 by testing H_0: $M_1(X) = M_2(X)$ for each $X \in \{X_1, \ldots, X_K\}$. Label the resulting p-values $p_1, \ldots, p_K$. The idea is to test the global hypothesis that (4) is true for every $k = 1, \ldots, K$ using some function of these K p-values. Perhaps the best-known method for testing some global hypothesis based on p-values is a technique derived by Fisher (1932). But Zaykin et al. (2002) note that the ordinary Fisher product test loses power in cases where there are a few large p-values. They suggest using instead a truncated product method (TPM), which is based on the test statistic

$$W = \prod p_k^{I(p_k \le \tau)}$$

where I is the indicator function (cf. Li & Siegmund, 2015). Setting $\tau = 1$ yields Fisher's method, but Zaykin et al. suggest using $\tau = .05$. Zaykin et al. derive the null distribution of W when all K tests are independent. But the K tests performed here are not independent simply because $P_j(X_k) \cap P_j(X_l)$, $k \neq l$, is not necessarily empty. If this dependence among the tests is ignored when computing a critical value for W, control over the Type I error probability is poor. For the dependent case, Zaykin et al. suggest using a bootstrap method, but this results in relatively high execution time for the situation at hand making this approach difficult to study via simulations. Consequently, an alternative approach was used: Momentarily assume normality and homoscedasticity with the goal of determining the α quantile of W, say w, in which case (4) is rejected at the α level if $W \le w$. Then study the impact of non-normality and heteroscedasticity via simulations.

The critical value w is determined via simulations using (2) with $M_j(X) \equiv 0$ and ε_j having a standard normal distribution. More precisely, for each j, (Y_{ij}, X_{ij}) ($i = 1, \ldots, n_j$; $j = 1,2$) are generated from a trivariate normal distribution where all correlations are zero. Then W is computed and this process is repeated say B times yielding $W_1, \ldots, W_B$. Put these B values in ascending order yielding $W_{(1)} \le \ldots \le W_{(B)}$. Then w is estimated to be $W_{(k)}$, where k is αB rounded to the nearest integer. Here, $B = 4000$ is used. Increasing the correlation to .5 had almost no impact on the estimated critical value.

One of many alternative methods is to use instead the test statistic

$$\bar{Q} = \frac{1}{K}\sum p_k$$

Wilcox (2016) found that this alternative test statistic performed relatively well, in terms of power, under a shift in location model. Now reject the global hypothesis if $\bar{q} \le q_\alpha$, the α quantile of $\bar{Q}$, which again is determined via simulations in the same manner as the critical value w. So rejecting indicates that one or more of the hypotheses given by (4) are false, but details about which ones are lacking.

Method M_3

The following strategy, called method M_3, is suggested for dealing with the limitation of method M_2. First, choose covariate points as done by method M_2. Based on this process for choosing covariate points, determine p_α, the α quantile of the distribution of the minimum p-value returned by method M_2. This is done via simulations in essentially the same manner used by method M_2. The only difference from method M_2 is that W and $\bar{Q}$ are replaced by $\tilde{p}$ = min(p_1, ..., p_K). So for a simulation based on B replications yielding $\tilde{p}_1, \ldots, \tilde{p}_B$, p_α is estimated with $\tilde{p}_{(k)}$, where k is the same as in method M_2 and $\tilde{p}_{(1)} \le \cdots \le \tilde{p}_{(B)}$ are the $\tilde{p}$ values written in ascending order. Then make a decision about whether $M_1(X)$ is larger than $M_2(X)$ for any covariate point for which the corresponding p-value is less than or equal to p_α. Otherwise, no decision is made. So method M_3 has the potential of providing more detail about where the regression surfaces differ. But of course there is the issue of how well it performs when dealing with non-normality, heteroscedasticity and curvature, which is examined via simulations in the next section. And another issue is the impact on power compared to method M_1.

Table 1. Some estimates of p_α, α = .05

n	p_α	*n*	p_α	*n*	p_α
50	0.00458	80	0.00248	400	0.00131
55	0.00320	100	0.00186	500	0.00135
60	0.00282	200	0.00142	600	0.00108
70	0.00259	300	0.00142	800	0.00096

Estimates of p_α, when $n_1 = n_2 = n$ are informative. Table 1 shows estimates for values of n ranging between 50 and 800 when α = .05. So the estimates appear to be converging to zero, but at an extremely slow rate. Consider, for example $n = 100$, in which case fifty hypotheses are tested. As indicated by Table 1, p_α is estimated to be .00186. Using the Bonferroni method instead, each hypothesis would be tested at the .0005 level, which is even less than the estimate of pα when using M_3 with n = 800.

Simulation Results

As is evident, an issue is the impact on the Type I error probability when dealing with non-normal distributions as well as situations where there is an association with the covariate variables. Simulations were used to address this issue with $n_1 = n_2 = 50$. Smaller sample sizes, such as $n_1 = n_2 = 30$, routinely result in situations where no covariate values can be found where comparisons can be made. That is, $N_1(X) < 12$ or $N_2(X) < 12$ for all $X \in \{X_1, \ldots, X_K\}$.

Estimated Type I error probabilities were based on 4000 replications. Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, values for the error term ε_j in (3) were generated from one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution, then by definition

$$V = \frac{\exp(gZ)-1}{g}\exp\left(hZ^2/2\right), \text{ if } g > 0$$
$$V = Z\exp\left(hZ^2/2\right), \text{ if } g = 0$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). The X_{ij} values were generated from a bivariate normal distribution with correlation equal to zero. Increasing this correlation to .5 altered the estimates of the Type I error probability by only a few units in third decimal place, so for brevity they are not reported.

Table 2. Some properties of the *g*-and-*h* distribution.

g	*h*	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Two types of regression surfaces were considered. The first deals with the situation where $Y = \lambda(X)\varepsilon$, which is labeled S_1. The second, labeled S_2, is $Y = X^2 + \lambda(X)\varepsilon$. Three choices for $\lambda(X)$ were considered: $\lambda(X_i) \equiv 1$ (VP_1), $\lambda(X_i) = |X_{i1}| + 1$ (VP_2) and $\lambda(X_i) = 1/(|X_{i1}| + 1)$ (VP_3). Estimated Type I error probabilities are reported in Table 3. Although the seriousness of a Type I error depends on the situation, Bradley (1978) suggested as a general guide, when testing at the .05 level, the actual level should be between .025 and .075. He goes on to suggest that ideally the actual level should be between .045 and .055. As can be seen, the estimates satisfy his first criterion, and nearly all of them satisfy his more stringent criterion.

Table 3. Estimated Type I error probabilities when testing at the $\alpha = .05$ level, $n_1 = n_2 = 50$

g	*h*	*S*	VP_1	VP_2	VP_3
0.000	0.000	1.000	0.050	0.052	0.050
0.000	0.000	2.000	0.056	0.050	0.048
0.000	0.200	1.000	0.046	0.039	0.049
0.000	0.200	2.000	0.048	0.050	0.053
0.200	0.000	1.000	0.052	0.050	0.044
0.200	0.000	2.000	0.054	0.048	0.050
0.200	0.200	1.000	0.051	0.048	0.048
0.200	0.200	2.000	0.055	0.040	0.044

In some situations, method M_2 can have substantially higher power than method M_3, where power is taken to be the probability of detecting one or more true differences. Consider, for example, the situation where for the first group $Y = \varepsilon$ and for the second group $Y = \varepsilon + .5$, where ε has a standard normal distribution and both sample sizes are 50. Then method M_2 has power approximately .41 compared to .26 using method M_3. If instead $Y = X^2 + \varepsilon$ for the second group, now power is .79 for M_2 and .65 using M_3. That is, M_2 might offer a substantial gain in power among the situations considered here at the expense of providing virtually no details about where significant results are obtained. However, methods M_2 and M_3 are sensitive to different features among the p-values. The next section illustrates that situations are encountered where M_3 rejects in contrast to M_2.

To provide some sense of how methods M_3 and M_1 compare in terms of power, again consider the situation where for the first group $Y = \varepsilon$ and for the second group $Y = \varepsilon + .5$. With both sample sizes equal to 100, power was estimated to be .51 and .42 for M_3 and M_1, respectively. If instead $Y = X^2 + \varepsilon$ for the second group, now power is .52 for M_3 and .51 using M_1. If $Y = X + \varepsilon$ for the second group, now the corresponding power estimates are .62 and .55. So, for at least some situations, method M_3 has substantially higher power than method M_1 despite the substantially larger number of hypotheses that are tested.

Illustrations

Data from the Well Elderly 2 study (Clark et al., 2011; Jackson et al., 2009) are used to illustrate that the choice between M_2 and M_3 can make a practical difference. A general goal in the Well Elderly 2 study was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the impact of intervention on a measure of perceived physical health, which was measured with the RAND 36-item (SF36) Health Survey, a measure of self-perceived physical health and mental well-being (Hays et al., 1993; McHorney et al., 1993). Higher scores reflect greater perceived health and well-being. There were two covariates. The first is a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). The CESD (Radloff, 1977) is sensitive to change in depressive status over time and has been successfully used to assess ethnically diverse older people (Lewinsohn et al., 1988; Foley et al., 2002). Higher scores indicate a higher level of depressive symptoms. The other covariate was the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicated that measures of stress are associated with the CAR. (The CAR is taken to be the cortisol level upon awakening minus the level of cortisol after the participants were awake for about an hour.) The sample size for the control group was 187 and the sample size for the group that received intervention was 228.

Based on both methods M_1 and M_2, no significant differences were found when testing at the .05 level. Method M_1 used only three covariate points. In contrast, method M_3 finds nine significant results among the 74 covariate points that were used. They occur where the CAR is negative (cortisol increases after awakening) and CESD is relatively low. So despite the simulation results indicating that M_2 can have higher power than M_3, situations are encountered where M_3 rejects and M_2 does not. Figure 1 shows a plot of the difference in SF36 scores (SF36 scores for the experimental group minus SF36 scores for the control group) as a function of the covariate points that were used. As can be seen, the largest differences occur when CESD scores are low and the CAR is negative. That is, intervention appears to be most beneficial, in terms of perceived health, for participants for whom cortisol increases after awakening. This is particularly true for participants who have low measures of depressive symptoms.

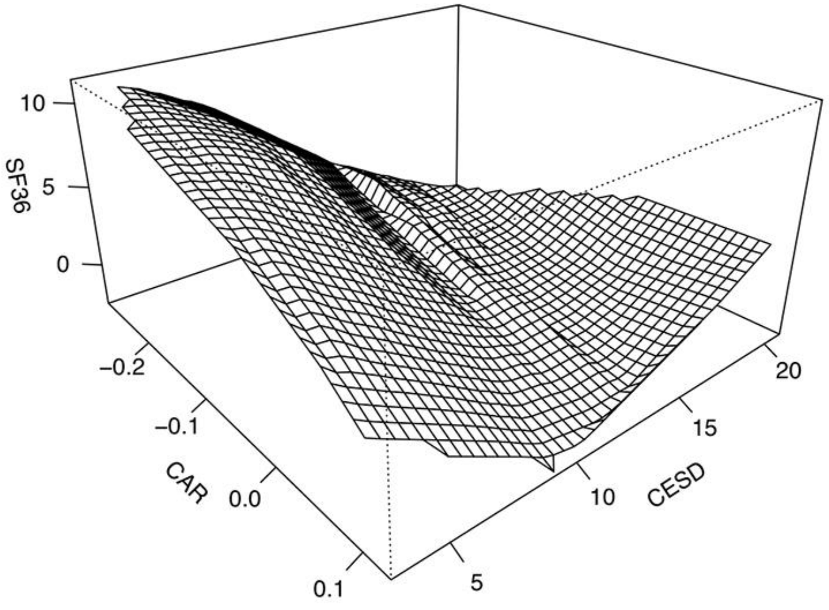

Figure 1. Regression surface predicting the typical difference in SF36 scores as a function of the CAR and CESD.

Conclusion

There are many variations of method M_3 that might have practical value. For example, some other measure of depth might be used or some alternative strategy for choosing the covariate points might offer an advantage. The main point is that based on simulations, all indications are that method M_3 controls the probability of one or more Type I errors very well. At least in some situations it offers a distinct power advantage over M_1 and no situation has been found where the reverse is true. There are situations where M_2 provides higher power than M_3, but at the cost of providing almost no details about where a significant difference occurs among the covariate points that were used.

In principle, methods M_1, M_2 and M_3 can be used when there is more than two covariates. But a general concern is the curse of dimensionality: neighborhoods with a fixed number of points become less local as the dimensions increase (Bellman, 1961). In practical terms, the expectation is that as the number of covariates increases, it becomes increasingly difficult to get an accurate estimate of the true regression surface.

References

Bellman, R. E. (1961). *Adaptive Control Processes*. Princeton, NJ: Princeton University Press.

Benjamini Y. & Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*, 1165–1188. doi: 10.1214/aos/1013699998

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Chida, Y. & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology*, *80*(3), 265–278. doi: 10.1016/j.biopsycho.2008.10.004

Clark, F., Jackson, J., Carlson, M., et al. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health*, *66*(9), 782–790. doi: 10.1136/jech.2009.099754

Clow, A., Thorn, L., Evans, P. & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress*, *7*(1), 29–37. doi: 10.1080/10253890410001667205

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics*, *20*, 1803–1827. doi: 10.1214/aos/1176348890

Eakman, A. M., Carlson, M. E. & Clark, F. A. (2010). The meaningful activity participation assessment: a measure of engagement in personally valued activities. *International Journal of Aging Human Development, 70*(4), 299–317. doi: 10.2190/ag.70.4.b

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer-Verlag. doi: 10.1007/b97679

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. New York: Marcel Dekker.

Fisher, R. (1932). *Statistical methods for research workers*. Oliver and Boyd, Edinburgh.

Foley K., Reed P., Mutran E., et al. (2002). Measurement adequacy of the CESD among a sample of older African Americans. *Psychiatric Research*, *109*(1), 61–69. doi: 10.1016/s0165-1781(01)00360-2

Fox, J. (2001). *Multiple and Generalized Nonparametric Regression*. Thousands Oaks, CA: Sage. doi: 10.4135/9781412985154

Györfi, L., Kohler, M., Krzyzk, A. & Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer Verlag. doi: 10.1007/b97848

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

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

Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

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

Heritier, S., Cantoni, E, Copt, S. & Victoria-Feser, M.-P. (2009). *Robust Methods in Biostatistics*. New York: Wiley. doi: 10.1002/9780470740538

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, pp. 461–515.

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, *75*(2), 383–386. doi: 10.1093/biomet/75.2.383

Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics* (2nd Ed). New York: Wiley. doi: 10.1002/9780470434697

Jackson, J., Mandel, D., Blanchard, J., et al. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials*, *6*(1), 90–101. doi: 10.1177/1740774508101191

Lewinsohn, P.M., Hoberman, H. M., Rosenbaum M. (1988). A prospective study of risk factors for unipolar depression. *Journal of Abnormal Psychology, 97*(3), 251–64. doi: 10.1037/0021-843x.97.3.251

Li, J. & Siegmund, D. (2015). Higher criticism: *p*-values and criticism. *Annals of Statistics*, *43*(3), 1323–1350. doi: 10.1214/15-aos1312

Long, J. S. & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54*(3), 217–224. doi: 10.1080/00031305.2000.10474549

Liu, R. Y., Parelius, J. M. & Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *Annals of Statistics, 27*(3), 783–858. 10.1214/aos/1018031260

Maronna, R. A., Martin, D. R. & Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. New York: Wiley. doi: 10.1002/0470010940

McHorney, C. A., Ware, J. E. & Raozek, A. E. (1993). The MOS 36-item Short-Form Health Survey (SF-36): II. Psychometric and clinical tests of validity in measuring physical and mental health constructs. *Medical Care*, *31*(3), 247–263. doi: 10.1097/00005650-199303000-00006

Ng, M. & Wilcox, R. R. (2011). A comparison of two-stage procedures for testing least- squares coefficients under heteroscedasticity. *British Journal of Mathematical and Statistical Psychology, 64*(2), 244-258. doi: 10.1348/000711010x508683

Radloff, L. (1977). The CESD scale: a self report depression scale for research in the general population. *Applied Psychological Measurement, 1*(3), 385–401. doi: 10.1177/014662167700100306

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika, 77*(3), 663–666. doi: 10.1093/biomet/77.3.663

Rosenberger, J. L. & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Understanding Robust and exploratory data analysis*. (pp. 297–336). New York: Wiley.

Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. doi: 10.1002/9781118165485

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd Ed). San Diego, CA: Academic Press.

Wilcox, R. R. (2016). ANCOVA: A heteroscedastic global test when there is curvature and two covariates. *Computational Statistics*, *31*(4), pp. 1593-1606. doi: 10.1007/s00180-015-0640-4

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

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., & Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology*, *22*(2), 170–185. doi: 10.1002/gepi.0042