DIGITALCOMMONS
@WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 2

3-6-2019

Robust ANCOVA, Curvature, and the Curse of Dimensionality

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

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2682
doi: 10.22237/jmasm/1551906370

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

Robust ANCOVA, Curvature, and the Curse of Dimensionality

Rand Wilcox
University of Southern California
Los Angeles, CA

There is a substantial collection of robust analysis of covariance (ANCOVA) methods that effectively deals with non-normality, unequal population slope parameters, outliers, and heteroscedasticity. Some are based on the usual linear model and others are based on smoothers (nonparametric regression estimators). However, extant results are limited to one or two covariates. A minor goal here is to extend a recently-proposed method, based on the usual linear model, to situations where there are up to six covariates. The usual linear model might provide a poor approximation of the true regression surface. The main goal is to suggest a method, based on a robust smoother, for dealing with curvature when there are three or four covariates. The results include perspectives on the curse of dimensionality. Perspectives on the use of a linear model versus a smoother are given.

Keywords: Smoothers, Multiple covariates, Trimmed means, Yuen's method, Heteroscedasticity

Introduction

Consider comparing two independent groups in a manner that takes into account p covariates. Let $\mathbf{X}_j$ be a vector of p covariates associated with the j^{th} group (j = 1, 2), and let Y_j be some outcome of interest. Let $\mathrm{M}_j(\mathbf{X})$ be some conditional measure of location associated with Y_j given $\mathbf{X} = (\mathbf{X}_1,\ldots, \mathbf{X}_p)$, where $\mathrm{M}_j(\mathbf{X})$ is some unknown function. The goal is to test

$$H_0 : \mathrm{M}_1(\mathbf{X}) = \mathrm{M}_2(\mathbf{X}) \tag{1}$$

doi: 10.22237/jmasm/1551906370 | Accepted: August 25, 2017; Published: March 6, 2019.
Correspondence: Rand Wilcox, rwilcox@usc.edu

Rand Wilcox is a Professor in the Department of Psychology, University of Southern California. His primary interests are robust and nonparametric statistical methods.

for a collection of covariate points in a manner that controls the probability of one or more Type I errors.

A common approach is to assume that, for the j^{th} group ($j = 1, 2$),

$$Y_j = \beta_{0j} + \beta_{1j} X_{j1} + \ldots + \beta_{pj} X_{jp} + \lambda\left(X_{j1}, \ldots, X_{jp}\right)\varepsilon_j \tag{2}$$

where $X_{j1}, \ldots, X_{jp}$ are the p covariates associated with the j^{th} group, $\beta_{0j}, \ldots, \beta_{pj}$ are unknown parameters, $\lambda(X_{j1}, \ldots, X_{jp})$ is some unknown function that models heteroscedasticity, and $\boldsymbol{\varepsilon}$ is some appropriate error term. The classic analysis of covariance method (ANCOVA) assumes that $\beta_{k1} = \beta_{k2}$ ($k = 1, \ldots, p$), $\lambda(X_{j1}, \ldots, X_{jp}) \equiv 1$ (within group homoscedasticity), and that ε_j has a normal distribution with mean zero and unknown variance σ_j^2 with $\sigma_1^2 = \sigma_2^2$ (between group homoscedasticity), in which case the goal is to test H_0: $\beta_{01} = \beta_{02}$. Moreover, least squares regression is used to estimate the unknown parameters.

There are several serious concerns with this classic method. First, least squares regression is not robust (e.g., Staudte & Sheather, 1990; Maronna, Martin, & Yohai, 2006; Heritier, Cantoni, Copt, & Victoria-Feser, 2010; Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber & Ronchetti, 2009; Wilcox, 2017a). Second, violating either of the two homoscedasticity assumptions can negatively impact both the control over the probability of a Type I error and power. Third, non-normality can negatively impact control over the probability of a Type I error and power as well. Fourth, outliers can destroy power and they can yield a highly misleading indication of the association within each group. A fifth limitation is that the slope parameters are assumed to be identical. Sixth, the linear model given by (2) might poorly approximate the true regression surface. There might be curvature that is poorly modeled by this linear model. There is now a substantial collection of techniques aimed at dealing with all of these concerns (e.g., Wilcox, 2017a, chapter 12).

Presumably, the linear model given by (2) provides an adequate approximation of the regression surface in some situations. But there is considerable evidence that often this is not the case (e.g., Hastie & Tibsherani, 1990; Wilcox, 2017a, b). Moreover, there are indications that, as the number of covariates increases, curvature becomes an increasing concern. One strategy is to include terms in (2) having the form X_{kj}^{a} for some choice for a, but it is known that this approach can be unsatisfactory even when $p = 1$ (e.g., Wilcox, 2017a). Numerous nonparametric regression estimators, generally known as smoothers, have been derived with the goal of dealing with curvature in a reasonably flexible manner. From a robustness point of view, the running interval smoother has proven to have considerable practical value. However, in terms of ANCOVA, extant results are limited to $p = 1$ or 2 covariates. The main goal here is to suggest a method for dealing with ANCOVA via the running interval smoother with the focus on $p = 3$ or 4 covariates. As will be seen, the method used here differs from the methods in Wilcox (2017a) in a manner to be described.

A method in Wilcox (2017c), based on the linear model given by (2), is readily extended to more than two covariates. Another goal is to report results on this alternative approach and to provide some sense of its relative merits. The method does not assume homoscedasticity and it does not assume that the regression slopes for each group are identical. A practical issue is how much is gained or lost when using the running interval smoother instead. The paper also comments of the relative merits of using certain diagnostic tools aimed at justifying the use of (2).

The Running Interval Smoother and Yuen's Method

Let (Y_{ij}, $\mathbf{X}_{ij}$) ($i = 1,\ldots, n_j$; $j = 1, 2$) be a random sample from the j^{th} group, where $\mathbf{X}_{ij}$ is a vector of p covariate values. Roughly, for the j^{th} group, the running interval smoother determines a subset of the $\mathbf{X}_{ij}$ vectors that are close to $\mathbf{X}$, then a measure of location is computed based on the corresponding Y_{ij} values, which yields an estimate of $\mathrm{M}_j(\mathbf{X})$. Here, the distance of $\mathbf{X}$ from each $\mathbf{X}_{ij}$ is based on a robust analog of Mahalanobis distance. To elaborate, let $\mathbf{S}$ be some covariance matrix. Then the distance between $\mathbf{X}$ and $\mathbf{X}_{ij}$ is

$$d_{ij} = \left(\mathbf{X} - \mathbf{X}_{ij}\right)' \mathbf{S}^{-1} \left(\mathbf{X} - \mathbf{X}_{ij}\right), \tag{3}$$

where $\mathbf{S}$ is taken to be the minimum covariance determinant estimator (e.g., Wilcox, 2017a, section 6.3.2). There are many other robust measures of covariance as well as robust measures of the distance of a point that are not based on some robust covariance matrix (Wilcox, 2017a, chapter 6). Perhaps one of these alternative choices offers a practical advantage for the situation at hand, but this issue goes beyond the scope of this paper.

Regarding the measure of location, here the focus is on the 20% trimmed mean. 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 $0 \leq \gamma < 0.5$, 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. A 20% trimmed mean corresponds to $\gamma = 0.2$, which has good efficiency relative to the sample mean under normality (Rosenberger & Gasko, 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 substantially reduces concerns about the impact of skewed distributions on the probability of a Type I error (e.g., Wilcox, 2017a). This is not to suggest that 20% trimming is always the optimal choice; clearly this is not the case. It is a reasonable choice among the many robust estimators that might be used.

For some constant f, generally known as the span, let $\mathrm{I}_j(\mathbf{X}) = \{i : d_{ij} \leq f\}$. That is, $\mathrm{I}_j(\mathbf{X})$ indexes the points that are close to $\mathbf{X}$. Then the estimate of $\mathrm{M}_j(\mathbf{X})$, $\hat{\mathrm{M}}_j(\mathbf{X})$, is the trimmed mean based on the Y_{ij} such that $i \in \mathrm{I}_j(\mathbf{X})$. For p = 1 or 2, a good choice for the span is often f = 0.8 or 1. This is not always the case, but these two values appear to perform reasonably well in general. Here, f = 1 is assumed unless stated otherwise. (The final section of this paper comments further on the choice for the span.) The main issue here is whether the ANCOVA method in the next section performs reasonably well in terms of controlling the probability of a Type I error.

Another issue is the so-called curse of dimensionality: neighborhoods with a fixed number of points become less local as the dimensions increase (Bellman, 1961). In practical terms, as p increases, what is the impact on the cardinality of $\mathrm{I}_j(\mathbf{X})$ for a given choice for the span, f? Results related to this issue are described in a later section in conjunction with the ANCOVA method described below.

Consider the goal of testing H_0: $\mu_{t1} = \mu_{t2}$, the hypothesis that two independent groups have identical trimmed means. For notational convenience, the method is described when ignoring the covariates. Derived by Yuen (1974), it is applied as follows: First, Winsorize the Y_{ij} values. That is, compute

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

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

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

Yuen's test statistic is

$$T_{\text{y}} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{d_1 + d_2}}$$

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

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

where

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

Description of the ANCOVA Method

Let $\mathrm{N}_j(\mathbf{X})$ be the cardinality of the set $\mathrm{I}_j(\mathbf{X})$. The basic idea is that if both $\mathrm{N}_1(\mathbf{X})$ and $\mathrm{N}_2(\mathbf{X})$ are reasonably large, Yuen's (1974) method for comparing trimmed means, based on the Y_{ij} such that $i \in \mathrm{I}_j(\mathbf{X})$, will generally provide reasonably adequate control over the Type I error probability when $\gamma = 0.2$. Following Wilcox (2017a), $\mathrm{N}_1(\mathbf{X})$ and $\mathrm{N}_2(\mathbf{X})$ are considered reasonably large if both are greater than or equal to 12. But there remains the issue of choosing which covariate points to use. If there are covariate points that have a particular substantive interest, and if the number of such points is relatively small, one can simply use the method in Wilcox (2017a, section 7.4.1) to control the probability of one or more Type I errors. But in various situations, such as an exploratory study, it might not be obvious which covariate points to use. Several strategies for choosing the covariate points are described in Wilcox (2017a). For example, determine the deepest half among the cloud of covariate points for the first group, which was the approach in Wilcox (2017c). For each such $\mathbf{X}$, if both $\mathrm{N}_1(\mathbf{X})$ and $\mathrm{N}_2(\mathbf{X})$ are greater than or equal to 12, perform Yuen's test. However, the details are not provided because here a different strategy is used: perform Yuen's test for each $\mathbf{X}_{ij}$ such that both $\mathrm{N}_1(\mathbf{X}_{i1})$ and $\mathrm{N}_2(\mathbf{X}_{i2})$ are greater than or equal to 12.

Consider how to control family wise Type I error rate (FWE), meaning the probability of one or more Type I errors. Let C be the number of covariate points such that both $\mathrm{N}_1(\mathbf{X}_{i1})$ and $\mathrm{N}_1(\mathbf{X}_{i2})$ are greater than or equal to 12. So, C reflects the number of hypotheses to be tested. If the goal is to perform the C tests so that the probability of one or more Type I errors is approximately α, a simple strategy is to reject the null hypothesis if $|T_\mathrm{y}| \geq q$, where q is the $1 - \alpha$ quantile of Studentized maximum modulus distribution. However, when the number of tests is relatively large, this approach becomes too conservative due to the strong association among the C tests. The probability of one or more Type I errors can be substantially smaller than the nominal α level, which in turn can negatively impact power.

Results in Wilcox (2017c) suggest how to proceed when $C > 25$. Let p_c ($c = 1, \ldots, C$) be the p-value associated with the c^{th} test and let $p_\mathrm{m} = \min(p_1, \ldots, p_C)$. The basic idea is to determine the α quantile of p_m, p_a, when all C hypotheses are true and when there is no association between Y and each of the p covariates. So, the probability of rejecting one or more hypotheses, when all C hypotheses are true, is $1 - \alpha$. That is, if any hypothesis is rejected when $p_c \leq p_a$, FWE will be α. But this leaves open the issue of well FWE is controlled when there is an association, an issue that is studied via simulations below.

To be more precise, consider the case $p = 3$. The α quantile of p_m was estimated as follows: Given p and $n = n_1 = n_2$, data were generated via (2) when all of the regression parameters are zero, the covariate values are generated from standard normal distributions all having correlation zero, $\lambda(X_{j1}, \ldots, X_{jp}) \equiv 1$, and when $\boldsymbol{\varepsilon}$ has a standard normal distribution. Then p_m was determined among the C

tests that were performed. This process was repeated 2000 times yielding 2000 p_{m} values. Next, a quantile regression smoother (Wilcox, 2017a, section 11.5.6) was used to estimate the regression line for predicting p_{α}, the α quantile of the distribution of p_{m}, given C, when $\alpha = 0.05$ (the method is based on the running interval smoother used in conjunction with the quantile estimator derived by Harrell & Davis, 1982). That is, p_a is the critical p-value when testing at the 0.05 level. The result suggested using a linear model for estimating p_a, given C, when $25 < C \leq 100$ and a different linear model when $C > 100$. This was done via the quantile regression estimator derived by Koenker and Bassett (1978). The results suggested estimating p_a, when $\alpha = 0.05$, with $0.0806452604/C - 0.0002461736$ when $25 < C \leq 100$. For $C > 100$, use 6.586286e-02/C + 4.137143e-05. In effect, this approach improves upon an approach based on the Studentized maximum modulus distribution by taking advantage of the strong association among the tests that are performed. This will be called method SM henceforth.

To add perspective, it is noted that with $p = 3$, $f = 1$, and $n = 50$, it was estimated that there is only a 0.003 probability that one or more tests would be performed. That is, due the curse of dimensionality, both $\mathrm{N}_1(\mathbf{X}_{i1})$ and $\mathrm{N}_1(\mathbf{X}_{i2})$ are typically less than 12. In practical terms, sample sizes greater than 50 are needed when dealing with curvature via the method used here. Increasing n to 80, this probability was estimated to be 0.612, and for $n = 150$ the estimate was 0.9995. For $n = 100$ the values of C ranged between zero and 27, with a median value of 9. For $n = 150$ they ranged between 18 and 68 with a median value of 47. With $p = 4$ and $n = 150$, the probability of performing one or more tests was estimated to be 0.23 with C ranging between 0 and 8. So without a fairly large sample size, large portions of the regression surfaces cannot be compared, which might result in missing important differences.

A Method Based on the Linear Model

The method above is readily modified for the situation where the linear model given by (2) is assumed to be true. In essence, the method described here is a generalization of the method in Wilcox (2017a, section 12.1.3), which is focused on $p = 2$ covariates only.

The method begins by pooling the covariate points for both groups and determining the deepest half of these points. There are various ways this might be done. Here projection distances are used. Roughly, projection distances are computed as follows: First determine the center of the data cloud. The marginal medians are used here, but there are several other robust location estimators that might be used (e.g., Wilcox, 2017a, section 6.3). Next, project all of the covariate points onto the line connecting the k^{th} covariate point and the center of data cloud. Then for each of the projected points, compute its distance from the center. This process is repeated for each k, and the projection distance of i^{th} point is taken to be its maximum distance among all of the projections. For a detailed description of the calculations, see Wilcox (2017a, section 6.2.5). If D_i is the distance of the i^{th} covariate point from the center of data cloud, its depth is taken to be $1 / (D_i + 1)$. Here, the R function pdepth, stored in the R package WRS, is used to compute projection depths. So here, the deepest half of the pooled covariate points to the center of the data cloud is used.

Consider the goal of testing (1). Compute some robust estimate of the regression parameters in (2), which yields an estimate of $\mathrm{M}_j(\mathbf{X})$, say $\tilde{\mathrm{M}}_j(\mathbf{X})$, for any $\mathbf{X}$ of interest. Here, the robust MM-estimator (Yohai, 1987) is used. The standard error of $\tilde{\mathrm{M}}_j(\mathbf{X})$ is estimated using a basic bootstrap method. For fixed j generate a bootstrap sample by resampling with replacement n_j vectors from $(Y_{ij}, \mathbf{X}_{ij})$. Based on this bootstrap sample, estimate $\mathrm{M}_j(\mathbf{X})$ yielding say $\tilde{\mathrm{M}}_j^*(\mathbf{X})$. Repeat this process B times yielding $\tilde{\mathrm{M}}_{jb}^*(\mathbf{X})$ ($b = 1,\ldots, B$). Then an estimate of the squared standard error of $\tilde{\mathrm{M}}_j(\mathbf{X})$ is

$$S^2 = \frac{1}{B-1}\sum\left(\tilde{\mathrm{M}}_{jb}^*(\mathbf{X}) - \bar{\mathrm{M}}_j^*(\mathbf{X})\right)^2$$

where $\bar{\mathrm{M}}_j^*(\mathbf{X}) = \sum \tilde{\mathrm{M}}_{jb}^*(\mathbf{X}) \big/ B$. Here, $B = 100$, which seems to suffice in a range of situations (e.g., Wilcox, 2017a). An appropriate test statistic for testing (1) is

$$V = \frac{\tilde{\mathrm{M}}_1(\mathbf{X}) - \tilde{\mathrm{M}}_2(\mathbf{X})}{S}$$

Simulations revealed a limitation associated with S: with small sample sizes, it is severely biased. More precisely, the actual standard error can be substantially smaller than indicated by S. For $p = 2$ covariates the bias becomes negligible when the sample size is at least 50. For $p = 3$, 4, 5, and 6 covariates, a sample size of 100 or more is required.

Let D denote the number of unique points among the deepest half of the covariate points. If all $n_1 + n_2$ covariate points are unique, then $D = (n_1 + n_2) / 2$. There remains the issue of controlling the probability of one or more Type I errors among all D tests that are performed. This is done by proceeding in a manner similar to the approach in the previous section. Momentarily focus on the case where all of the slope parameters are zero, there is homoscedasticity, and the error term has a standard normal distribution. Generate data for both groups and compute a p-value assuming that V has a standard normal distribution. This is done for each covariate point of interest yielding D p-values. Next, determine the smallest p-value. This process was repeated 2000 times for sample sizes 50, 100, 500, and 800; and for $p = 2$, 3, 4, 5, and 6. This yields an estimate of the null distribution of the smallest p-value among the D tests that are performed.

For $p = 2$, the critical 0.05 p-value changes very little as the sample size increases, provided the smallest sample size is at least 50. In particular, if (1) is rejected whenever a p-value is less than or equal to 0.00615847, the probability of one or more Type I errors, when all D hypotheses are true, is approximately 0.05. For sample sizes less than 50, this probability can be substantially smaller than 0.05 due to the bias associated with S. For $p = 2$, 3, 4, 5, and 6, the critical 0.05 p-values are 0.002856423, 0.00196, 0.001960793, and 0.001120947, respectively. Now the probability of one or more Type I errors is approximately 0.05 provided the smallest sample size is at least 100. More generally, these critical p-values appear to be approximately correct when the number of tests is greater than 25. Otherwise, using a critical value based on the Studentized maximum modulus distribution, with infinite degrees of freedom, seems preferable. Again, with smaller sample sizes, the probability of one or more Type I errors can be substantially smaller than 0.05 due to the bias associated with S. The method in this section is called method LIN henceforth.

Simulation Results

Simulations were used as a partial check on the ability of the methods in the preceding sections to control the family wise error (FWE) rate when testing at the 0.05 level. First the focus is on method SM and then results for method LIN are reported. Four types of distributions are considered for the error term: normal, symmetric and heavy-tailed (roughly meaning that outliers tend to be common), asymmetric and relatively light-tailed, and asymmetric and relatively heavy-tailed. More specifically, data are generated from g-and-h distributions (Hoaglin, 1985). If Z has a standard normal distribution, then by definition

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

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 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Hoaglin (1985) summarizes additional properties of the *g*-and-*h* distributions. As for the independent variables, they were generated from a bivariate normal distribution with correlation zero or 0.6.

Data were generated from the model

$$Y = \sum_{j=1}^{p} X_j^a + \lambda\left(X_{j1}, \ldots, X_{jp}\right)\varepsilon \tag{4}$$

where $a = 1$ or 2, $p = 3$ or 4, and ε has one of the *g*-and-*h* distributions shown in Table 1. Two choices for $\lambda(X_{j1},\ldots, X_{jp})$ were used: $\lambda(X_{j1},\ldots, X_{jp}) \equiv 1$ (homoscedasticity) and $\lambda(X_{j1},\ldots, X_{jp}) = |X_{j1} + X_{j2}| + 1$ (heteroscedasticity). The results are reported in Table 2 based on 2000 replications, where the column headed by HOM are the results when there is homoscedasticity and HET indicates heteroscedasticity. The common correlation is zero; results when the common correlation is 0.6 did not reveal any additional insights.

Although the seriousness of a Type I error can depend on the situation, Bradley (1978) has suggested that as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. As indicated in Table 2, all of the estimated Type I error probabilities fall in this range.

Table 1. 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

Table 2. Estimates of FWE when using SM and testing at the 0.05 level

g	*h*	*n*	*p*	*a*	**HOM**	**HET**
0.0	0.0	80	3	1	0.044	0.041
0.0	0.0	80	3	2	0.047	0.041
0.0	0.2	80	3	1	0.042	0.038
0.0	0.2	80	3	2	0.043	0.036
0.2	0.0	80	3	1	0.046	0.040
0.2	0.0	80	3	2	0.048	0.040
0.2	0.2	80	3	1	0.042	0.037
0.2	0.2	80	3	2	0.043	0.034
0.0	0.0	150	3	1	0.051	0.053
0.0	0.0	150	3	2	0.050	0.046
0.0	0.2	150	3	1	0.045	0.040
0.0	0.2	150	3	2	0.038	0.035
0.2	0.0	150	3	1	0.054	0.054
0.2	0.0	150	3	2	0.046	0.046
0.2	0.2	150	3	1	0.042	0.036
0.2	0.2	150	3	2	0.040	0.034
0.0	0.0	200	4	1	0.044	0.039
0.0	0.0	200	4	2	0.047	0.039
0.0	0.2	200	4	1	0.039	0.031
0.0	0.2	200	4	2	0.037	0.028
0.2	0.0	200	4	1	0.040	0.038
0.2	0.0	200	4	2	0.043	0.035
0.2	0.2	200	4	1	0.034	0.028
0.2	0.2	200	4	2	0.037	0.029

Table 3. Estimates of FWE when using LIN and testing at the 0.05 level

g	*h*	*n*	*p*	**HOM**	**HET**
0.0	0.0	100	3	0.046	0.051
0.0	0.2	100	3	0.033	0.041
0.2	0.0	100	3	0.040	0.043
0.2	0.2	100	3	0.030	0.039
0.0	0.0	200	3	0.042	0.044
0.0	0.2	200	3	0.041	0.036
0.2	0.0	200	3	0.043	0.045
0.2	0.2	200	3	0.040	0.046
0.0	0.0	100	6	0.048	0.040
0.0	0.2	100	6	0.033	0.034
0.2	0.0	100	6	0.050	0.038
0.2	0.2	100	6	0.033	0.028
0.0	0.0	200	6	0.052	0.057
0.0	0.2	200	6	0.046	0.050
0.2	0.0	200	6	0.054	0.054
0.2	0.2	200	6	0.036	0.048

Table 4. Estimated power assuming a linear model

g	*h*	*n*	*p*	SM	LIN
0.0	0.0	100	3	0.206	0.490
0.0	0.2	100	3	0.170	0.438
0.2	0.0	100	3	0.202	0.485
0.2	0.2	100	3	0.169	0.440
0.0	0.0	200	4	0.208	0.804
0.0	0.2	200	4	0.168	0.786
0.2	0.0	200	4	0.204	0.804
0.2	0.2	200	4	0.165	0.785

Now consider method LIN in the previous section. Simulations were run for the same distributions used in connection with method SM. Estimates of the FWE are shown in Table 3 for $p = 2$ and 6 when all correlations among the covariates are zero. Again, the column headed by HOM indicates results when the error term is homoscedastic and HET indicates the heteroscedastic case. Very similar results were obtained when all correlations are equal to 0.6. As can be seen, all indications are that method LIN performs reasonably well based on Bradley's criterion.

Next, the power of methods SM and LIN are compared when the linear model is true and there is a shift in location. More precisely, for the first group data were generated based on (2) where all of the slope parameters are equal to one, the intercept is zero, there is homoscedasticity, the error term has a standard normal distribution, and the independent variables have a multivariate normal distribution with all correlations equal to zero. Data for the second group were generated in exactly the same manner only the intercept was taken to be δ. Table 4 shows the estimated power when $\delta = 0.5$. Not surprisingly, LIN always has more power. Table 4 illustrates that the increase in power can be substantial.

When the linear model is wrong, the reverse can happen. Consider, for example, a situation where $p = 4$, $n = 200$, all of the slopes for the first group are zero, otherwise the situation is the same as in Table 4. Now imagine that for the second group, $Y = X_1^2 + \varepsilon - 1$. Method SM was estimated to have power 0.490 versus 0.174 for method LIN. If instead for the second group $Y = X_1^2 + X_2^2 + \varepsilon - 2$, the power estimates for methods SM and LIN are now 0.902 and 0.317, respectively.

A natural strategy is to consider diagnostic tools for choosing between SM and LIN. For example, one might test the hypothesis that the linear model given by (2) is correct. A robust method for accomplishing this goal is described in Wilcox (2017a, section 11.6.1). But it is unclear when this method will have enough power to detect a situation where the departure from the linear model is enough to create practical concerns. Another approach is to use the diagnostic strategy derived by Berk and Booth (1995) for detecting curvature. If curvature does not appear to be an issue, this would seem to suggest using method LIN. Again, it is unclear when this approach is able to detect a situation where power, for example, will be higher using SM rather than LIN. A third possibility is to plot the predicted value of *Y* using the linear model versus the predicted values using method SM. If the plotted points appear to be centered around a line having slope one and intercept zero, this would seem to suggest that there is little or no advantage to using method SM. The next section illustrates that all three of these strategies can be unsatisfactory in the sense that they appear to support the use of method LIN, yet method SM yields significant results while method LIN does not.

Illustration

Methods SM and LIN are illustrated with data from the Well Elderly 2 study (Clark et al., 2012), which was generally designed to assess the effectiveness of an intervention program aimed at improving the physical and emotional wellbeing of older adults. Depressive symptoms (CESD) are compared to a control group taking into account three covariates: measures of life (LSIZ) satisfaction, meaningful activities (MAPA), and stress. The sample size for the control group is 232 and 141 for the experimental group.

Consider method SM. It performs comparisons for a total of 85 covariate points, five of which were significant. For the control group, the median values for LSIZ, MAPA, and stress were 18, 32, and 4, respectively. The corresponding lower quartiles were 14, 28, and 2; and the upper quartiles were 21, 36, and 6. For the experimental group, the medians were 18, 32, and 4. The lower quartiles were 14, 28, and 2; and the upper quartiles were 22, 36, and 6. The covariate points where a significant difference was found were (19, 38, 1), (17, 34, 1), (18, 35, 1), (22, 40, 0), and (21, 38, 0). So significant differences are found when LSIZ and MAPA are relatively high among the participants, and when stress is relatively low.

Now consider the linear model given by (2). The next goal is to illustrate some issues and concerns again using the Well Elderly 2 data. For all of the analyses reported here, leverage points (covariate points flagged as outliers) are removed. Testing the hypothesis that the model given by (2) is correct, when the MM-estimator is used, the *p*-value for the control group is 0.63 and for the experimental group it is 0.19. The R function lintest was used; see Wilcox (2017a) for details. But as previously noted, it is unclear whether power is sufficiently high to detect a departure from the linear model that creates practical concerns. If the MM-estimator is replaced by the Theil (1950) and Sen (1968) estimator, for the experimental group there is now a significant result at the 0.05 level: the *p*-value is 0.046. So, the choice of which regression estimator is used can matter. However, one might argue that even if there is some departure from the linear model, perhaps it is of little or no consequence. As a partial check on this possibility, Figure 1 shows the smooth for predicting $\tilde{M}_2(\mathbf{X})$, the fitted values based on the linear model, versus $\hat{M}_2(\mathbf{X})$, the fitted values based on the running interval smoother. The solid line is the predicted value of $\hat{M}_2(\mathbf{X})$ given $\tilde{M}_2(\mathbf{X})$. The dashed line is a line having slope one and intercept zero. So, the plot suggests that there is little practical difference between the two methods. Partial residual plots (Berk & Booth, 1995) also suggest that assuming a liner model provides a reasonably accurate approximation of the regression surface. The R function prplot, described in Wilcox (2017a, section 14.4.7), was used.

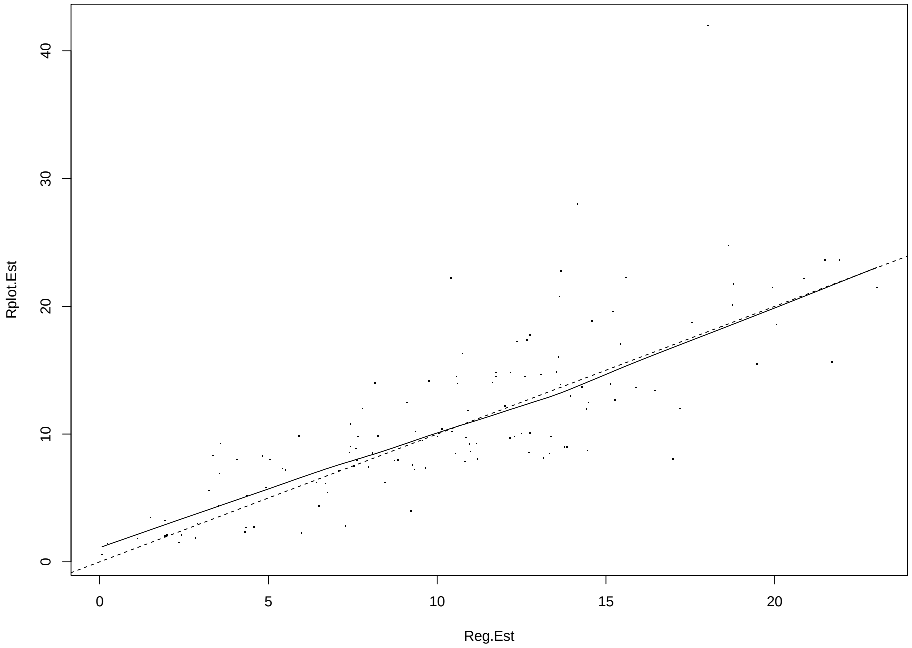

Figure 1. The *x*-axis corresponds to $\tilde{M}_2(\mathbf{X})$, the estimate of CESD based on the linear model; the *y*-axis corresponds to given $\hat{M}_2(\mathbf{X})$, the estimate of CESD based on the running interval smoother; the solid line is the regression line for predicting $\tilde{M}_2(\mathbf{X})$ given $\hat{M}_2(\mathbf{X})$; the dashed line has slope one and intercept zero; so, the plot provides some indication of the extent the two regression estimators give similar results.

Based on the results just described, an argument might be made that assuming a linear model is reasonable with the hope that this will yield more power. However, for the situation at hand, method LIN does not reject for any of 161 covariate points used. If the analysis is limited to the same covariate points used by method SM, again no significant results are obtained.

Conclusion

An obvious advantage of method LIN is that it is less restrictive in terms of the number of covariates and the sample sizes that can be accommodated. And there is some possibility that it can have substantially more power when the linear model is true. But this comes at a price. In essence, LIN ignores any issues related to the curse of dimensionality. Moreover, even if diagnostic tools suggest that the linear model provides a reasonable approximation of the regression surface, it is possible for method SM to reject when method LIN does not, as was illustrated in the final section. For the moment, the suggestion is to use SM if possible. If there are indications that a linear model is reasonable, also use method LIN, as might be done in an exploratory study.

Another issue is the span used by method SM. While the choice used here appears to be reasonable in general, if curvature is sufficiently severe, a smaller choice for the span can be required to get a reasonable approximation of the regression surface. Diagnostic tools for detecting such situations are in need of further development. The point here is that in the context of ANCOVA, reducing the span exacerbates concerns associated with the curse of dimensionality; much larger sample sizes are needed than those indicated here.

The R function ancdetM4 applies method SM and ancJNPVAL applies method LIN. Both of these functions have been added to the file Rallfun-v35, which can be downloaded from https://dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm. Both functions are also available in the R package WRS, which is located at https://github.com/nicebread/WRS.

References

Bellman, R. E. (1961). *Adaptive control processes: A guided tour*. Princeton, NJ: Princeton University Press.

Berk, K. N., & Booth, D. E. (1995). Seeing a curve in multiple regression. *Technometrics, 37*(4), 385-398. doi: 10.1080/00401706.1995.10484372

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

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., … Azen, S. P. (2012). 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 & Community Health, 66*(9), 782-790. doi: 10.1136/jech.2009.099754

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

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

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. New York, NY: Chapman and Hall. doi: 10.1201/9780203753781

Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2010). *Robust methods in biostatistics*. New York, NY: Wiley. doi: 10.1002/9780470740538

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. C. Hoaglin, F. Mosteller, & J. W. Tukey (Eds.), *Exploring data tables, trends, and shapes* (pp. 461-511). New York, NY: Wiley. doi: 10.1002/9781118150702.ch11

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

Koenker, R., & Bassett, G., Jr. (1978). *Regression quantiles*. Econometrica, 46(1), 33-50. doi: 10.2307/1913643

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

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

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association, 63*(324), 1379-1389. doi: 10.1080/01621459.1968.10480934

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

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. I. *Indagationes mathematicae, 12*(3), 85-91.

Wilcox, R. R. (2017a). *Introduction to robust estimation and hypothesis testing* (4th ed.). San Diego, CA: Academic Press.

Wilcox, R. (2017b). *Modern statistics for the social and behavioral sciences: A practical introduction* (2nd ed.). New York, NY: Chapman & Hall/CRC. doi: 10.1201/9781315154480

Wilcox, R. R. (2017c). Robust ANCOVA: Heteroscedastic confidence bands that have some specified simultaneous probability coverage. *Journal of Data Science*, *15*(2), 313-328. Retrieved from http://www.jds-online.com/file_download/607/2017_4-8.pdf