Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 3

5-1-2006

ANCOVA: A Robust Omnibus Test Based On Selected Design Points

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

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 14-21

1538 – 9472/06/$95.00

ANCOVA: A Robust Omnibus Test Based On Selected Design Points

Rand R. Wilcox
Dept of Psychology
University of Southern California

Many robust analogs of the classic analysis of covariance method have been proposed. One approach, when comparing two independent groups, uses selected design points and then compares the groups at each design point using some robust method for comparing measures of location. So, if K design points are of interest, K tests are performed. There are rather obvious ways of performing, instead, an omnibus test that for all K points, no differences between the groups exist. One of the main results here is that several variations of these methods can perform very poorly in simulations. An alternative approach, based in part on the usual sample median, is suggested and found to perform reasonably well in simulations. It is noted that when using other robust measures of location, the method can be unsatisfactory.

Key words: ANCOVA, bootstrap methods, measures of depth, smoothers

Introduction

The analysis of covariance (ANCOVA) problem is to compare two independent groups based on some outcome of interest, Y , in a manner that takes into account some covariate, X. A classic and well-known approach assumes that the error term of the usual linear regression model is homoscedastic and has a normal distribution, the regression lines associated with each group are parallel, and the variances associated with the error terms for each group are assumed to be identical. More formally, if for the jth group *(j* = 1, 2), then there are n_j randomly sampled pairs of observations, say (X_{ij}, Y_{ij}), $i = 1, . . . n_j$, the classic assumption is that for the jth group,

$$Y_{ij} = \beta X_{ij} + \beta_{oj} + \varepsilon_{ij} \qquad (1)$$

where ε_{ij} has variance $\sigma_j^2, \sigma_1^2 = \sigma_2^2$, and ε_{ij} is independent of X_{ij}. So by implication, for each group, the conditional variance of Y , given X, does not vary with X, and each group has the same slope.

It is known that violating one or more of these assumptions can result in serious practical problems. Concerns about the robustness of the method date back to at least Atiqullah (1964) who concluded that non-normality is a practical problem. Another obvious concern is the assumption that the regression lines are parallel. There are several robust methods for testing this assumption (e.g., Wilcox, 2003, 2005), but it remains unclear when such tests have enough power to detect situations where having non-parallel lines is a practical concern. Yet another concern about equation (1) is the assumption that the association between Y and X is linear.

Rand R. Wilcox is Professor of Psychology at the University of Southern California, Los Angeles. Email: rwilcox@usc.edu.

Of course, in some situations this is a reasonable approximation, but this is not always the case. Many alternative methods have been derived that eliminate the assumption that the association is linear (e.g. Bowman & Young, 1996; Delgado, 1993; Dette & Neumeyer, 2001; Hall, Huber, & Speckman, 1997; Kulasekera, 1995; Kulasekera & Wang, 1997; Munk & Dette, 1998; Neumeyer & Dette, 2003; Young & Bowman, 1995; Wilcox, 2003). However, some of these methods require homoscedasticity and for most there are few if any simulation results that support their use with small to moderate sample sizes.

A simple and very flexible approach to ANCOVA is described in Wilcox (2003, section 14.8). It allows the regression lines to be non-linear, it allows heteroscedasticity, it performs well in simulations, and in the event standard assumptions are met, all indications are that it has nearly the same amount of power as the classic ANCOVA method (e.g., Wilcox, 2005, p. 526). Roughly, the method is based on multiple comparisons. Examination of the method suggests a simple and rather obvious approach to performing an omnibus test instead. But results reported here make it clear that several variations of this approach perform very poorly in simulations. (Details are given later in the article). The main result in this article is that an alternative approach, based in part on the usual sample median and the depth of the null vector in a bootstrap cloud, nearly eliminates this problem. The main exception is a situation where, simultaneously, the conditional distribution of Y is discrete, skewed, and the possible values for Y are relatively small in number.

Considered and Discarded Methods

It helps to describe the first general method that was considered and discarded and then suggest a related approach that gives more satisfactory results. It is assumed that for the jth group, Y and X are related through some unknown function, m_j. More formally, it is assumed that

$$Y_{ij} = m_j(X_{ij}) + \varepsilon_{ij}$$

where ε_{ij} has a median of zero, variance σ_{ij}^2, and is independent of X_{ij}. Let $m_j(x)$ be the population median of Y for the jth group, given that the covariate of the jth group is $X_j = x$. (Comments on using other location estimators are given later in the article). Let $x_1, ..., x_K$ be K values of X that are of interest. The method in Wilcox (2003, section 14.8) includes as a special case the problem of testing

$$H_0 : m_1(x_k) = m_2(x_k), k = 1, ..., K,$$

for each k. That is, K tests are to be performed. Let $\delta(x_k) = m_1(x_k) - m_2(x_k)$. The goal here is to test

$$H_0 : \delta(x_1) = \cdots = \delta(x_K) = 0 \qquad (2)$$

Here, it is assumed that K = 5 and that the choices for $x_1, ..., x_5$ are made empirically in a manner about to be described. Of course, it is not being suggested that other choices for the design points or K are inappropriate. For example, a researcher might have interest in K specific design points, rather than points determined as is done here. The idea is to provide a data-driven method for checking whether the regression lines differ, paying particular attention to design points where valid inferences about the medians of the Y values can be made.

The choice of the five design points stems in part from what is called a running interval smoother. To describe the details, attention is temporarily focused on a single group of subjects. The basic strategy is to find all X_i values close to x and estimate *m(x)* with the median of the corresponding Y values. The method begins by computing the median absolute deviation statistic:

$$MAD = median\{|X_1 - M|, ..., |X_n - M|\},$$

where M is the usual sample median of the X values. Let MADN = MAD/.6745. The only reason for rescaling MAD is that under normality, MADN estimates σ. This rescaling helps describe the running interval smoother in terms of familiar concepts, but ultimately it is not important. Then X_i is said to be close to x if

$$| X_i - x | \leq f \times MADN,$$

where f is some constant, called the span. Here, following Wilcox (2003), f = 1 is used. Let $\overline{m}_j = \Sigma m_j(x_k)/K$. A seemingly natural alternative to (2) is to test

$$H_0 : \overline{m}_1 = \overline{m}_2 \qquad (3)$$

That is, view the problem in the context of a 2 by K ANOVA and test the hypothesis that there is no main effect for the first factor. Many robust methods for testing this hypothesis have been proposed (Wilcox, 2005), which include various bootstrap techniques. But when checking the ability of this approach to control the probability of a Type I error for the problem at hand, poor results were obtained in situations described later in the article. Included were non-bootstrap methods for 20% trimmed means and medians (Wilcox, 2003, sections 10.3 & 10.5) plus bootstrap variations of these methods described in Wilcox (2005). In particular, it was found that in some situations, when testing at the .05 level, the actual Type I error probability was estimated to exceed .2.

Description of the Recommended Method

The one method that performed well in simulations is based on testing (2) rather than (3). The general strategy is to generate bootstrap samples, yielding bootstrap estimates of δ_k, and then determine how deeply the null vector is nested within this bootstrap cloud. Two approaches to measuring the depth of the null vector are considered. General theoretical results related to this approach are reported in Liu and Singh (1997).

To elaborate, momentarily assume that the xk values have been chosen and let Y_{ijk} $(i = 1..., n_{jk}; k = 1,..., K)$ be the Y_{ij} values such that

$$| X_{ij} - x_k | \leq f \times MADN. \qquad (4)$$

For fixed k and j, generate a bootstrap sample by randomly sampling with replacement n_{jk} values from Y_{ijk} yielding $Y^*_{ijk}, (i = 1,..., n_{jk})$. Let M^*_{jk} be the usual sample median based on the Y^*_{ijk} values and let $\delta^*_k = M^*_{1k} - M^*_{2k}$. Repeat this process B times yielding $\delta^*_{bk}, b = 1,..., B$. So, there are B vectors of bootstrap δ^*_{bk} values, each vector having length K. Then roughly, the null hypothesis is rejected depending on how deeply the null vector $(0,...,0)$ is nested within this bootstrap cloud.

The problem of choosing the x_k values is approached as follows. Let $N_j(x)$ be the number of points in the jth group that are considered close to x based on (4). For notational convenience, assume that for fixed j, the X_{ij} values are in ascending order. That is, $X_{1j} \leq \cdots \leq X_{njJ}$. The regression lines are said to be comparable at x if simultaneously $N_j(x) \geq 12$ for both j = 1 and 2. The value 12 is chosen simply to reflect a sample of points large enough so as to expect reasonable control over the probability of a Type I error, but obviously some other (larger) value could be used if desired.

Suppose x_1 is taken to be the smallest X_{i1} value for which the regression lines are comparable. That is, search the first group for the smallest X_{i1} such that $N_1(X_{i1}) \geq 12$. If $N_2(X_{il}) \geq 12$, the two regression lines are considered comparable at X_{i1} and $x_1 = X_{i1}$ is set. If $N_2(x_{il}) < 12$, consider the next largest X_{i1} value and continue until it is simultaneously true that $N_1(X_{i1}) \geq 12$ and $N2(Xi1) \geq 12$. $K = 5$ is used, but again some other value is certainly reasonable. Let x_5 be the largest X_{i1} value in the first group for which the regression lines are comparable. That is, x_5 is the largest X_{i1} value such that $N_1(x_5) \geq 12$ and $N_2(x_5) \geq 12$. Let i_5 be the corresponding value of i. The other three design points are chosen as follows. Let $i_3 = (i_1 + i_5)/2$, $i_2 = i_1 + i_3/2$, and $i_4 = (i_3 + i_5)/2$. Round i_2, i_3, and i_4 down to the nearest integer and set $x_2 = X_{i_2 1}$, $x_3 = X_{i_3 1}$, and $x_4 = X_{i_4 1}$.

There are various ways of measuring how deeply a point is nested within a multivariate cloud of data (e.g., Liu & Singh, 1997, Wilcox, 2005). The simplest is based on Mahalanobis distances and is the first of the two methods considered here. However, the most obvious estimate of the covariance matrix associated with the bootstrap vectors is not used. Rather, it is estimated with

$$s_{km} = \frac{1}{B-1}\sum_{b=1}^{B}(\delta^*_{bk} - \delta_k)(\delta^*_{bm} - \delta_m).$$

That is, for fixed k, rather than use $\Sigma\delta^*_{bk}/B$ as the estimate of the center of the bootstrap cloud, use δ_k instead. Put another way, there is no need to estimate the center of the bootstrap cloud, it is already known and given by the vector $(\delta_1, ..., \delta_K)$. Indeed, if it is estimated with $\Sigma\delta^*_{bk}/B$, control over the probability of a Type I error deteriorates, consistent with a variety of other methods surveyed by Wilcox (2005). Let $S = (s_{km})$ be the corresponding covariance matrix, in which case the distance of the bth bootstrap vector from the center is given by

$$d_b = \sqrt{(\delta^*_{b1} - \delta_1, ..., \delta^*_{bK} - \delta_K)S^{-1}(\delta^*_{b1} - \delta_1, ..., \delta^*_{bK} - \delta_K)'}.$$

Let

$$D = \sqrt{(\delta_1 - 0, ..., \delta_K - 0)S^{-1}(\delta_1 - 0, ..., \delta_K - 0)'},$$

which is the distance of the null vector from the center of the bootstrap cloud. The (generalized) p-value is

$$\hat{p}^* = \frac{1}{B}\Sigma I(D \leq d_b),$$

where $I(D \leq d_b) = 1$ if $D \leq d_b$ and $I(D \leq d_b) = 0$ if $D > d_b$. This will be called method M.

The second method considered here for measuring the depth of a point in the bootstrap cloud is a projection-type method given in Wilcox (2005, section 6.2.5); it represents a slight variation of a method discussed by Donoho and Gasko (1992) and has been found to perform well in connection with other methods described in Wilcox (2005). The computational details are relegated to an appendix. This will be called method P.

A Simulation Study

Simulations were used to assess the small-sample properties of the method just described. Observations were generated according to the models

$$Y = \varepsilon$$

$$Y = X + \varepsilon$$

and

$$Y = X^2 + \varepsilon,$$

where X has a standard normal distribution and ε has one of four g-and-h distributions (Hoaglin, 1985), which contain the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$W = \begin{cases} \dfrac{\exp(gZ)-1}{g}\exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

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

g	h	k_1	k_2
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	1.75	8.9
0.2	0.2	2.81	155.99

has a g-and-h distribution, where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0.0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and a symmetric distribution with heavy tails ($g = h = 0.2$). In Table 1, the theoretical skewness and kurtosis for each distribution is considered. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

A general concern about methods aimed at comparing population medians, based on the usual sample median, is that for discrete data where tied values can occur, control over the probability of a Type I error can be poor. This is the case when using the method proposed by Bonett and Price (2002) as well as a related method in Wilcox (2003, section 8.7.1). In a paper submitted for publication, the author has found that certain bootstrap methods correct this problem while others do not. The main point here is that considering discrete distributions where tied values are likely is crucial for the problem at hand. Accordingly, additional simulations were run by generating ε from a beta-binomial distribution:

$$P(X = x) = \frac{B(m-x+r, x+s)}{(m+1)B(m-x+1, x+1)B(r,s)},$$

where B is the complete beta function. Here m = 10, 12 and 20 were considered. With m = 12, for example, the possible values for X are the integers $0,1,\ldots,12$. The values for r and s were taken to be $r = s = 4$, as well as $r = 1$ and $r = 9$. For $r = s = 4$ the distribution is bell-shaped and symmetric with mean $m/2$. In Figure 1, the probability function when $r = 1$, $s = 9$ and $m = 12$ is exhibited.

In Table 2, the estimated probability of a Type I error when testing at the .05 level and $n_1 = n_2 = 40$ is exhibited. The estimates are based on 1,000 replications with $B = 600$. (From Robey & Barcikowski, (1992), 1,000 replications is sufficient from a power point of view. More specifically, if the hypothesis that the actual Type I error rate is .05 is tested, and if power is to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required.) The results for $Y = X + \varepsilon$ did not reveal any new insights, and so for brevity they are not reported. To get some idea of the effect of homoscedasticity, additional simulations were run where values in the first group were multiplied by $\sigma_1 = 4$. The g-and-h distribution has a median of zero, so the null hypothesis remains true. For the beta-binomial distributions, the data were shifted to have a median of zero before multiplying by $\sigma_1 = 4$. The top portion of Table 2 are the results when there is homoscedasticity $(\sigma_1 = 1)$.

Figure 1: The beta-binomial probability function with $m = 12$, $r = 1$ and $s = 9$

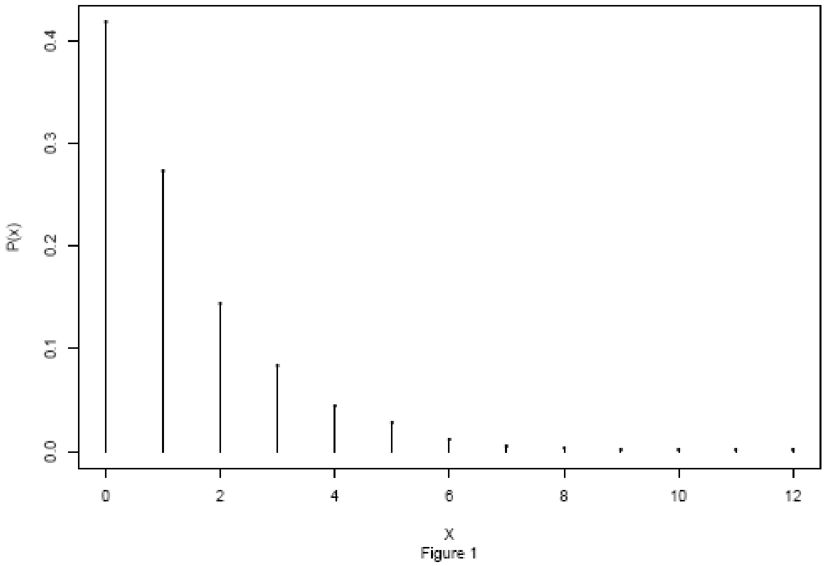

Table 2: Estimated Type I error probabilities

		$\sigma_1 = 1$			
ϵ		$Y = \epsilon$		$Y = X_1^2 + \epsilon$	
g	h	P	M	P	M
0.0	0.0	.064	.059	.052	.054
0.0	0.2	.039	.065	.047	.057
0.2	0.0	.061	.073	.050	.061
0.2	0.2	.041	.064	.048	.054
m	r, s				
12	1, 9	.464	.071	.053	.066
20	1, 9	.428	.027	.055	.066
10	4, 4	.152	.058	.063	.068
		$\sigma_1 = 4$			
g	h	P	M	P	M
0.0	0.0	.048	.088	.061	.089
0.0	0.2	.042	.076	.052	.076
0.2	0.0	.047	.089	.060	.087
0.2	0.2	.038	.077	.053	.076
m	r, s				
12	1, 9	.267	.142	.096	.146
20	1, 9	.133	.081	.077	.101
10	4, 4	.062	.055	.055	.073

First, consider the homoscedastic case with continuous g-and-h distributions. Both methods P and M perform reasonably well. To avoid an estimated Type I error probability greater than .07, method P is preferable. Under heteroscedasticity, method M can be unsatisfactory, with estimates exceeding .08, while again method P gives fairly satisfactory results. But when tied values occur, method P can be disastrous and should not be used. Method M now performs well under homoscedasticity $(\sigma_1 = 1)$, but under heteroscedasticity, it breaks down as well with estimates exceeding .1.

All simulations were repeated with $n_1 = n_2 = 60$, no new insights were found, so the results are not reported.

Conclusion

A positive result is that when tied values occur with probability zero, method P performs fairly well in terms of Type I errors, even when there is heteroscedasticity. However, when tied values are likely, it can be unsatisfactory. If tied values are likely and there is homoscedasticity, method M performs reasonably well, but it can break down when there is heteroscedasicity. So a possible argument in favor of method M is that when the (conditional) distributions of Y do not differ, it provides good control over the probability of a Type I error. But a negative feature is that it is sensitive to more than one feature of the data. That is, it does not isolate the reason for rejecting, which could be due to differences between medians or heteroscedasticity.

Some additional simulations were run with m = 2 0,r = 2 and s = 9. The ability of method P to control the probability of a Type I error improved substantially versus the situation where r = 1, but the estimated probability of a Type I error for the model $Y = \varepsilon$ was .099. So it seems that some tied values can probably be tolerated when using method P, but it is difficult to know when this is the case.

A criticism of the sample median is that under normality, or when sampling from a light-tailed distribution, it is relatively inefficient. By trimming less, say 20%, good efficiency is obtained under normality and some protection against low efficiency due to heavy-tailed distributions is obtained. (Note that the usual sample median belongs to the class of trimmed means with the maximum amount of trimming.) However, replacing the usual sample median with a 20% trimmed mean, the methods studied here are unsatisfactory in terms of estimated Type I errors, at least for the situations considered. Consideration was given to estimating the population median with the Harrell and Davis (1982) estimator with the goal of achieving better efficiency under normality, but again control over the probability of a Type I error was no longer satisfactory.

References

Atiqullah, M. (1964). The robustness of the covariance analysis of a one-way classification. *Biometrika, 51,* 365–372.

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.

Bowman, A. & Young, S. (1996). Graphical comparison of nonparametric curves. *Applied Statistics, 45,* 83–98.

Delgado, M. A. (1993). Testing the equality of nonparametric regression curves. *Statistics and Probability Letters, 17,* 199–204.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics, 27,* 1012–1040.

Dette, H. & Neumeyer, N. (2001). Nonparametric analysis of covariance. *Annals of Statistics, 29,* 1361–1400.

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.

Hall, P., Huber, C., & Speckman, P. L. (1997). Covariate-matched one-sided tests for the difference between functional means. *Journal of the American Statistical Association, 92,* 1074–1083.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69,* 635–640.

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.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *Journal of the American Statistical Association, 90*, 1085–1093.

Kulasekera, K. B. & Wang, J. (1997). Smoothing parameter selection for power optimality in testing of regression curves. *Journal of the American Statistical Association, 92*, 500–511.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on data depth and bootstrap. *Journal of the American Statistical Association, 92*, 266–277,

Munk, A. & Dette, H. (1998). Nonparametric comparison of several regression functions: Exact and asymptotic theory. *Annals of Statistics, 26*, 2339–2368.

Neumeyer, N. & Dette, H. (2003). Nonparametric comparison of regression curves: An empirical process approach. *Annals of Statistics, 31*, 880–920.

Robey, R. R. & Barcikowski, R. S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology, 45*, 283–288.

Young, S. G. & Bowman, A. W. (1995). Nonparametric analysis of covariance. *Biometrics, 51*, 920–931.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques Testing*. San Diego CA: Academic Press.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing* (2nd Ed). San Diego CA: Academic Press.

Appendix

For notational convenience, projection distance is described in terms of a sample of *n* vectors from some multivariate distribution. The sample is denoted by $X_i, i = 1,...,n$. Let ξ be some multivariate measure of location. Here, ξ is taken to be the W-estimator stemming from the minimum volume ellipsoid estimator. (For a detailed discussion of the minimum volume ellipsoid estimator, see Rousseeuw & Leroy, 1987). The outlier detection method in Rousseeuw and van Zomeren (1990) is applied, any points flagged as outliers are removed, and ξ is taken to be the mean of the remaining vectors. For any *i*, let

$$U_i = X_i - \xi \, ,$$

$$B_i = U_i U_i'$$
$$= \Sigma_{k=1}^{p} U_{ik}^2$$

and for any *j* let (j=1,…,n) let

$$W_{ij} = \sum_{k=1}^{p} U_{ik} U_{jk} \, ,$$

and

$$T_{ij} = \frac{W_{ij}}{B_i}(U_{i1},...,U_{ip}) \qquad (5)$$

The distance between ξ and the projection of X_j (when projecting onto the line connecting X_i and ξ) is

$$V_{ij} = \| T_{ij} \| ,$$

where $\| T_{ij} \|$ is the Euclidean norm associated with the vector T_{ij}. Let

$$d_{ij} = \frac{V_{ij}}{q_2 - q_1} , \qquad (6)$$

where for fixed i, q_2 and q_1 are estimates of the upper and lower quartiles, respectively, of the V_{ij} values. (Here, the ideal fourths based on the values $V_{i1},...V_{in}$ were used; see, for example, Wilcox, 2004.) The projection distance associated with X_j say D_j, is the maximum value of d_{ij}, the maximum being taken over $i = 1,...,n$.