Journal of Data Science **11**(2013), 281-291

A Heteroscedastic Method for Comparing Regression Lines at Specified Design Points When Using a Robust Regression Estimator

Rand R. Wilcox
University of Southern California

Abstract: It is well known that the ordinary least squares (OLS) regression estimator is not robust. Many robust regression estimators have been proposed and inferential methods based on these estimators have been derived. However, for two independent groups, let $\theta_j(X)$ be some conditional measure of location for the jth group, given X, based on some robust regression estimator. An issue that has not been addressed is computing a $1-\alpha$ confidence interval for $\theta_1(X)-\theta_2(X)$ in a manner that allows both within group and between group hetereoscedasticity. The paper reports the finite sample properties of a simple method for accomplishing this goal. Simulations indicate that, in terms of controlling the probability of a Type I error, the method performs very well for a wide range of situations, even with a relatively small sample size. In principle, any robust regression estimator can be used. The simulations are focused primarily on the Theil-Sen estimator, but some results using Yohai's MM-estimator, as well as the Koenker and Bassett quantile regression estimator, are noted. Data from the Well Elderly II study, dealing with measures of meaningful activity using the cortisol awakening response as a covariate, are used to illustrate that the choice between an extant method based on a nonparametric regression estimator, and the method suggested here, can make a practical difference.

Key words: ANCOVA, bootstrap methods, Theil-Sen estimator, Well Elderly II study.

1. Introduction

For two independent groups, assume that for the jth group $(j = 1, 2)$ Y_j is some outcome variable of interest and X_j is some covariate such that

$$Y_j = \beta_{0j} + \beta_{1j}X_j + \lambda_j(X)\epsilon_j, \tag{1}$$

where β_{0j} and β_{1j} are unknown parameters and ϵ_j is a random variable having variance σ_j^2 and mean equal to zero. So based on (1),

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

is the some conditional measure of location for Y given X. Classic inferential methods based on (1) assume two types of homoscedasticity. The first is within group (WG) homoscedasticity, meaning that $\lambda_j(X) \equiv 1$ and the other is between group (BG) homoscedasticity, meaning that $\sigma_1^2 = \sigma_2^2$. And of course, these methods are based on the least squares regression estimator. It is well known, however, that the ordinary least squares (OLS) regression estimator is not robust (e.g., Huber and Ronchetti, 2009; Hampel *et al.*, 1986; Staudte and Sheather, 1990; Wilcox, 2012). Included among the concerns about OLS is that its efficiency can be relatively poor when the error term ϵ_j has a heavy-tailed distribution, particularly when there is WG heteroscedasticity (Wilcox, 2012, p. 515). Another concern is that even a single outlier can result in a distorted and misleading summary of the association among the bulk of the points. In addition, when there is WG heteroscedasticity, classic methods are using an invalid estimate of the relevant standard errors (e.g., Godfrey, 2006; Long and Ervin, 2000). Numerous robust regression estimators have been derived that are aimed at dealing with known concerns associated with OLS, and inferential methods based on these estimators have been developed as well (e.g., Heritier *et al.*, 2009; Maronna *et al.*, 2006; Wilcox, 2012). But evidently, there are no results on methods aimed at testing

$$H_0 : \theta_1(X) = \theta_2(X), \tag{2}$$

for some specified value for X, which allow both types of heteroscedasticity. The goal in this paper is to suggest a simple method for accomplishing this goal and to report simulation results on how well it performs when the sample sizes are relatively small.

A related goal is comparing the slopes and intercepts based on robust regression estimators, and such techniques have already been derived (e.g., Wilcox, 2012). Of course, if the regression lines are not parallel, this raises the issue of determining the range of X values for which there is a high degree of certainty that $\theta_1(X) < \theta_2(X)$, as well as a range of X values for which we can be reasonably certain that $\theta_1(X) > \theta_2(X)$. As is evident, testing (2) helps address these problems (cf. Johnson and Neyman, 1936; Wilcox, 1987).

Wilcox (2012, Section 11.11) summarizes a collection of methods aimed at testing (2) based in part on a nonparametric regression estimator. More precisely, a running interval smoother is used to estimate $\theta_j(X)$ that makes no assumptions about the parametric form of the regression lines. Let $(Y_{1j}, X_{1j}), \cdots, (Y_{n_j j}, X_{n_j j})$ be a random sample from some bivariate distribution corresponding to the jth group. Briefly, given X, the method determines the X_{ij} values that are close to X in terms of a robust measure of variation (the median absolute deviation). To elaborate, let

$$N_j(X) = \{i : |X_{ij} - X| \leq f_j \times \mathrm{MADN}_j\},$$

where MAD_j is the usual sample median based on $|X_{1j} - M_j|, \cdots, |X_{n_j j} - M_j|$, M_j is the median of $X_{1j}, \cdots, X_{n_j j}$ and MADN_j is $\text{MAD}_j/0.6745$. Under normality, $\text{MAD}_j/0.6745$ estimates the usual population standard deviation. So roughly, under normality, X is said to be close to X_{ij} if X is within f_j standard deviations of X_{ij}. The constant f_j, called a span, is chosen in a manner that provides a reasonably good approximation of the regression line even when there is curvature. (Using $f_j = 0.8$ or 1 usually gives good results.) The method estimates $\theta_j(X)$ with a trimmed mean applied to the corresponding Y_{ij} values, meaning the Y_{ij} values such that $i \in N_j(X)$. The hypothesis of equal trimmed means is tested with the method derived by Yuen (1974). Bootstrap methods are available as well. It is evident that if indeed there is curvature, this approach can have more power than any method that assumes the regression lines are straight. But simultaneously, if indeed the regression lines are reasonably straight, there is the concern that power might be reduced substantially compared to a method that assumes there is no curvature. The reason is that given X, the nonparametric method uses only those Y_{ij} values such that $i \in N_j(X)$ when testing (2). As illustrated in Section 3, the method proposed here can indeed provide a substantial increase in power compared to the nonparametric method just described, and it can make a difference in practice, as illustrated in Section 4.

2. Description of the Proposed Method

There are many robust regression estimators. Here the focus is on the Theil (1950) and Sen (1968) regression estimator, but this is not to suggest that it dominates all other regression estimators that might be used. Indeed, no single estimator dominates based on the various criteria used to compare estimators. But the Theil-Sen estimator performs relatively well in terms of handling outliers and it has good efficiency, particularly when there is heteroscedasticity.

For convenience, momentarily consider a single group and let $(Y_1, X_1), \cdots, (Y_n, X_n)$ be a random sample from some bivariate distribution. The regression estimator proposed by Theil (1950) was based on the strategy of finding a value for the slope, b_1, that makes Kendall's correlation tau between $Y_i - b_1 X_i$ and X_i (approximately) equal to zero. Sen (1968) showed that this is tantamount to the following method. For any $i < i'$, for which $X_i \neq X_{i'}$, let

$$S_{ii'} = \frac{Y_i - Y_{i'}}{X_i - X_{i'}}.$$

The Theil-Sen estimate of the slope is b_1, the median of all the slopes represented by $S_{ii'}$. The intercept is estimated with

$$M_y - b_1 M_x,$$

where M_y and M_x are the usual sample medians of the Y and X values, respectively. Dietz (1987) showed that the Theil-Sen estimator has an asymptotic breakdown point of 0.293, where roughly the breakdown point of an estimator refers to the proportion of points that must be altered to make it arbitrarily large or small. So about 29% of the points must be altered in order to make the Theil-Sen estimate of the slope (or intercept) arbitrarily large or small. In practical terms, it offers protection against the event of outliers completely distorting the nature of the association among the bulk of the points. In contrast, OLS has a breakdown point of $1/n$. That is, a single outlier can result in a highly misleading summary of the association. For results on the small-sample efficiency of the Theil-Sen estimator, see Dietz (1989), Talwar (1991) and Wilcox (1998).

Various strategies for testing (2) were considered that did not perform well in simulations. For brevity, attention is focused on the one method that was found to be reasonably satisfactory. First, generate a bootstrap sample from the jth group by randomly sampling n_j vectors of observations, with replacement, from $(X_{1j}, Y_{1j}), \cdots, (X_{n_j j}, Y_{n_j j})$ yielding $(X^*_{1j}, Y^*_{1j}), \cdots, (X^*_{n_j j}, Y^*_{n_j j})$. Compute the Theil-Sen estimate of the slope and intercept based on this bootstrap sample and label the results β^*_{1j} and β^*_{0j}, respectively. For X specified, let $\hat{Y}^*_j = \beta^*_{0j} + \beta^*_{1j}X$. Repeat this process B times yielding $\hat{Y}^*_{jb}$ $(b = 1, \cdots, B)$. Then, from basic principles (e.g., Efron and Tibshirani, 1997), an estimate of the squared standard error of $\hat{\theta}_j(X) = b_{0j} + b_{1j}X$ is

$$\hat{\tau}^2_j = \frac{1}{B-1} \sum (\hat{Y}^*_{jb} - \bar{Y}^*_j)^2, \tag{3}$$

where $\bar{Y}^*_j = \sum \hat{Y}^*_{jb}/B$. Here, $B = 100$ is used. Letting z be the $1-\alpha/2$ quantile of a standard normal distribution, an approximate $1-\alpha$ confidence interval for $\theta_1(X) - \theta_2(X)$ is

$$\hat{\theta}_1(X) - \hat{\theta}_2(X) \pm z\sqrt{\hat{\tau}^2_1 + \hat{\tau}^2_2}. \tag{4}$$

Note that the method is readily extended to the case of $p > 1$ covariates.

When testing (2) for two or more choices for X, an issue is controlling the familywise error (FWE) rate, meaning the probability of one more Type I errors. There are, of course, various strategies that might be used, such as some sequentially rejective method that improves on the Bonferroni method (e.g., Hochberg, 1988; Rom, 1990), or one might simply replace z in (4) with the $1-\alpha$ quantile of a C-variate Studentized maximum modulus distribution with infinite degrees of freedom, where C is the number of tests to be performed. A few results based on this last strategy are reported in Section 3 when C is relatively small. Hochberg's method was considered as well, it was found to be less satisfactory, so the details of the method are omitted. However, in the final section of the paper, it is argued that this issue is in need of further research.

3. Simulation Results

Simulations were used to study the small-sample properties of the proposed method. The sample sizes were $(n_1, n_2) = (20, 20)$, (20, 40), (40, 40) and (200, 200). Simulations with $n_1 = n_2 = 200$ were run as a partial check on the R function that was used. Estimated Type I error probabilities, $\hat{\alpha}$, were based on 2000 replications. Four types of distributions were used: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, the marginal distributions were taken to be one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. (The R function ghdist, in Wilcox, 2012, was used to generate observations from a g-and-h distribution.) If Z has a standard normal distribution, then

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

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

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

g	h	κ_1	κ_2
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

The intercept was taken to be $\beta_0 = 0$ and two choices for the slope were used: $\beta_1 = 0$ and $\beta_1 = 1$. Three choices for λ were used: $\lambda(X) = 1$, $\lambda(X) = |X| + 1$ and $\lambda(X) = 1/(|X| + 1)$. For convenience, these three choices are denoted by variance patterns (VP) 1, 2, and 3. As is evident, VP 1 corresponds to the usual homoscedasticity assumption. VP 2 is a situation where the conditional variance of Y is relatively large when X is close to the median of its distribution; for values of X far from the median of its distribution the conditional variance of Y is relatively small. VP 3 is the reverse of VP 2.

Table 2 summarizes the simulation results for $n_1 = n_2 = 20$, $\beta_1 = 0$ and $\alpha = 0.05$ when using the bootstrap method in conjunction with the Theil-Sen estimator. Results when $\beta_1 = 1$ are similar to those in Table 2, so for brevity they are not reported. The column headed by FWE is the estimated probability of one or more Type I errors when testing (2) at $X = -1$, 0 and 1. The column headed by $\hat{\alpha}$ is the estimated probability of a Type I error when $X = -1$. (Similar results are obtained for $X = 0$ and 1.) As can be seen, the highest estimated level is 0.052. The main difficulty is that in some situations, the estimate drops as low as 0.014 when testing a single hypothesis, and FWE is estimated to be as low as 0.011. This occurs for VP 3 and when sampling from a relatively heavy-tailed distribution ($h = 0.2$). With $(n_1, n_2) = (20, 40)$, the estimated levels are much closer to the nominal level. For example, for $g = h = 0.2$, the estimate of FWE for VP 3 is 0.026 and $\hat{\alpha} = 0.035$. Using Hochberg's method instead, the estimated level is smaller than those reported in Table 2. For example, under normality and homoscedasticity, FWE using Hochberg's method was estimated to 0.022.

Table 2: Estimates of α and FWE, $n_1 = n_2 = 20$

g	h	VP	$\hat{\alpha}$	FWE
0.0	0.0	1	0.042	0.039
0.0	0.0	2	0.052	0.045
0.0	0.0	3	0.029	0.019
0.0	0.2	1	0.035	0.027
0.0	0.2	2	0.039	0.037
0.0	0.2	3	0.014	0.012
0.2	0.0	1	0.041	0.037
0.2	0.0	2	0.045	0.042
0.2	0.0	3	0.030	0.017
0.2	0.2	1	0.031	0.026
0.2	0.2	2	0.041	0.036
0.2	0.2	3	0.016	0.011

Table 2 does not report any results on the nonparametric ANCOVA method because for $n_1 = n_2 = 20$, typically the method cannot be applied. The reason is that, given X, the cardinality of the set $N_1(X)$ or $N_2(X)$ can be too small. If, for example, there are only two points or less in $N_1(X)$ say, then performing Yuen's test cannot be applied. This problem tends to occur regardless of the value for the covariate X that is chosen, but with $n_1 = n_2 = 40$ it can be avoided.

As previously mentioned, it is evident that the nonparametric ANCOVA method in Wilcox (2012) can have more power than the method studied here due to curvature. A practical issue is whether the reverse ever happens. Consider the case of normality and homoscedasticity, $\beta_{12} = \beta_{02} = 0$, but $\beta_{12} = 0.5$ and $\beta_{02} = 0.65$. With $n_1 = n_2 = 40$, the method in Section 2 has power approximately equal to 0.82 for $X = -1$, and the nonparametric method has power 0.73.

4. An Illustration

In the Well Elderly II study by Jackson *et al.* (2009), a general goal was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. A portion of the study was aimed at understanding the impact of intervention on a measure of meaningful activities as measured by the Meaningful Activity Participation Assessment (MAPA) instrument (Eakman *et al.*, 2010). Extant studies (e.g., Clow *et al.*, 2004; Chida and Steptoe, 2009) indicate that measures of stress are associated with the cortisol awakening response (CAR), which is defined as the change in cortisol concentration that occurs during the first hour after waking from sleep. (CAR is taken to be the cortisol level after the participants were awake for about an hour minus the level of cortisol upon awakening.)

Here, MAPA scores for a control group are compared to a group that received intervention using CAR as a covariate. Figure 1 shows a scatterplot of the data and the Theil-Sen regression lines for the two groups, where the solid line corresponds to the control group. (Points indicated by 0 correspond to the control group. Also, leverage points were removed.) The sample sizes are $n_1 = 232$ and $n_2 = 141$. A running interval smoother suggests that the regression line when predicting MAPA with CAR is well approximated by a straight line and a test of the hypothesis that the regression lines are straight (using the R function lintest in Wilcox, 2012, Section 11.6.2) failed to reject at the 0.05 level. (The p-values were $p = 0.28$ for the control group and 0.34 for the intervention group). For the points $X = -0.32$ and 0.15, the p-values corresponding to the null hypothesis given by (2) are 0.048 and 0.044, respectively, suggesting that with a reasonably high degree of certainty, the lines cross somewhere inside the interval $(-0.32, 15)$. In contrast, using the nonparametric method based on the running interval smoother, the p-values are 0.13 and 0.10, respectively.

5. Concluding Remarks

It is evident that simulations cannot establish that a method performs well in all situations that might be encountered. However, all indications are that the method considered here performs well when the Theil-Sen estimator is used, except with a very small sample size, in which case the actual level can drop well below the nominal level. Moreover, the non-normal distributions and the types of heteroscedasticity used in the simulations would seem to cover a wide range of situations. Also, alternative bootstrap methods that were considered performed poorly. Consequently, the method studied here is recommended.

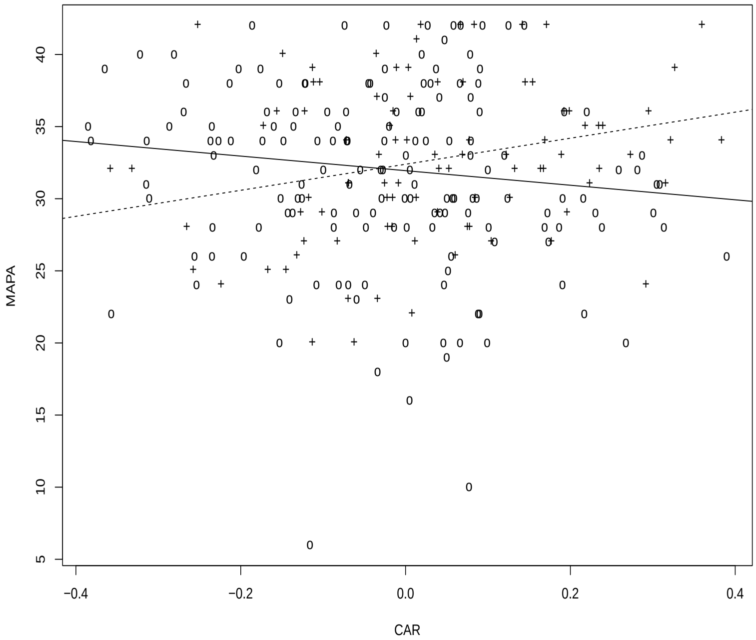

Figure 1: Regression lines for predicting MAPA with CAR. The solid line is the control group and the dashed line is the group that received intervention

A few simulations were run with the Theil-Sen estimator replaced by the robust MM-estimator derived by Yohai (1987) or the Koenker and Bassett (1978) quantile regression estimator. With $n_1 = n_2 = 20$, situations are encountered where these two alternative estimators cannot be computed based on a bootstrap sample. This problem did not occur with $n_1 = n_2 = 40$. With $n_1 = n_2 = 40$, it appears that control over the Type I error probability is similar to using the Theil-Sen estimator, with the actual levels being a bit lower. But a more extensive study is needed.

Although FWE was controlled reasonably well using the Studentized maximum modulus distribution, a speculation is that Hochberg's method will provide more power in some situations despite the results reported here. The reason is that if the covariate values are sufficiently similar, the p-values when testing (2) will be similar as well. Imagine, for example, that three hypotheses are tested and that all three p-values are equal to 0.049. Then by Hochberg's method, all three would be rejected at the 0.05 level, but using the Studentized maximum modulus distribution, none would be rejected.

The method studied here is readily generalized to more than one covariate. Some additional simulations were run with two predictors, X_1 and X_2, with $n_1 = n_2 = 40$, where now for the heteroscedastic case $\lambda(X_1, X_2) = |X_1| + 1$ and $\lambda(X_1, X_2) = 1/(|X_1| + 1)$ were used. For the distributions in Table 2, the estimated probability of a Type I error at $(X_1, X_2) = (0, 0)$ ranged between 0.021 and 0.038. But this issue is in need of further study.

A related goal is finding some global test of the hypothesis that the regression lines do not differ for any covariate value that might be chosen. That is, the goal is to test the hypothesis that simultaneously, $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$. There are methods for performing separate tests of

$$H_0 : \beta_{01} = \beta_{02},$$

and

$$H_0 : \beta_{11} = \beta_{12},$$

via a robust regression estimator that allows heteroscedasticity (e.g., Wilcox, 2012), but there are no results on a single global test of the hypothesis that both $\beta_{01} = \beta_{02}$ and $\beta_{11} = \beta_{12}$ are true. Methods for accomplishing this goal are under investigation.

Finally, the R function ancpar is available for applying the method in Section 2. It is contained in the R package Rallfun-v19, which is stored at http://college.usc.edu/labs/rwilcox/home.

References

Chida, Y. and Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: a systematic review and meta-analysis. *Biological Psychology* **80**, 265-278.

Clow, A., Thorn, L., Evans, P. and Hucklebridge, F. (2004). The awakening cortisol response: methodological issues and significance. *Stress* **7**, 29-37.

Dietz, E. J. (1987). A comparison of robust estimators in simple linear regression. *Communications in Statistics - Simulation and Computation* **16**, 1209-1227.

Dietz, E. J. (1989). Teaching regression in a nonparametric statistics course. *American Statistician* **43**, 35-40.

Eakman, A. M., Carlson, M. E. and Clark, F. A. (2010). The meaningful activity participation assessment: a measure of engagement in personally valued activities. *International Journal of Aging Human Development* **70**, 299-317.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

Godfrey, L. G. (2006). Tests for regression models with heteroskedasticity of unknown form. *Computational Statistics & Data Analysis* **50**, 2715-2733.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.

Heritier, S., Cantoni, E, Copt, S. and Victoria-Feser, M. P. (2009). *Robust Methods in Biostatistics.* Wiley, New York.

Hoaglin, D. C. (1985). *Summarizing Shape Numerically: The g-and-h Distribution.* In *Exploring Data Tables, Trends, and Shapes* (Edited by D. C. Hoaglin, F. Mosteller and J. W. Tukey), 461-514. Wiley, New York.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* **75**, 800-802.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*, 2nd edition. Wiley, New York.

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C. P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B. and Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials* **6**, 90-101.

Johnson, P. O. and Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs* **1**, 57-93.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33-50.

Long, J. S. and Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician* **54**, 217-224.

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

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika* **77**, 663-665.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association* **63**, 1379-1389.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* Wiley, New York.

Talwar, P. P. (1991). A simulation study of some non-parametric regression estimators. *Computational Statistics & Data Analysis* **15**, 309-327.

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

Wilcox, R. R. (1987). Pairwise comparisons of J independent regression lines over a finite interval, simultaneous pairwise comparison of their parameters, and the Johnson-Neyman technique. *British Journal of Mathematical and Statistical Psychology* **40**, 80-93.

Wilcox, R. R. (1998). A note on the Theil-Sen regression estimator when the regressor is random and the error term is heteroscedastic. *Biometrical Journal* **40**, 261-268.

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

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics* **15**, 642-656.

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika* **61**, 165-170.

Received August 25, 2012; accepted November 14, 2012.

Rand R. Wilcox
Department of Psychology
University of Southern California
Los Angeles, CA 90089-1061, USA
rwilcox@usc.edu