5-1-2007

On Flexible Tests of Independence and homoscedasticity

Rand R. Wilcox

University of Southern California, rwilcox@usc.edu

On Flexible Tests of Independence and Homoscedasticity

Rand R. Wilcox
University of Southern California

Consider the nonparametric regression model $Y = m(X) + \tau(X)\varepsilon$, where X and ε are independent random variables, ε has a mean of zero and variance σ^2, τ is some unknown function used to model heteroscedasticity, and $m(X)$ is an unknown function reflecting some conditional measure of location associated with Y, given X. Detecting dependence, by testing the hypothesis that $m(X)$ does not vary with X, has the potential of being more sensitive to a wider range of associations compared to using Pearson's correlation. This note has two goals. The first is to point out situations where a certain variation of an extant test of this hypothesis fails to control the probability of a Type I error, but another variation avoids this problem. The successful variation provides a new test of $H_0 : \tau(X) \equiv 1$, the hypothesis that the error term is homoscedastic, which has the potential of higher power versus a method recently studied by Wilcox (2006). The second goal is to report some simulation results on how this method performs.

Key words: Heteroscedasticity, smoothers, wild bootstrap, Winsorized correlations.

Introduction

Consider the nonparametric regression model $Y = m(X) + \tau(X)\varepsilon$, where X and ε are independent random variables, ε has a mean of zero and variance σ^2, τ is some unknown

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology at the University of Southern California. He is the author of seven textbooks on statistics, the most recent of which is *Introduction to Robust Estimation and Hypothesis Testing* (2005, 2nd Ed., San Diego, CA: Academic Press).

Pearson's correlation between X and Y. A concern about this approach is that it limits the function used to model heteroscedasticity, and $m(X)$ is an unknown function reflecting some conditional measure of location associated with Y, given X. Typically, it is assumed that $m(X) = \beta_0 + \beta_1 X$, where β_0 and β_1 are the unknown slope and intercept, and of course a common approach toward establishing an association is testing $H_0 : \rho = 0$, where ρ is types of associations between X and Y that can be detected. For example, there are many types of curvilinear associations between X and Y for which ρ will be close to zero, which in turn can mean relatively low power when

testing $H_0 : \rho = 0$. Also, heteroscedasticity can affect power when using the usual Student's T test of this hypothesis (e.g., Wilcox, 2003), roughly because the wrong standard error is being used. In some cases, heteroscedasticity might increase power, but the reverse can happen as well. Perhaps more importantly, when Student's T rejects, there is uncertainty whether the main reason is due to heteroscedasticity or because $m(X)$ varies with X.

A test of $H_0 : \rho = 0$ that allows heteroscedasticity is given in Wilcox (2003, section 7.3.2), but again, curvilinear associations might be missed. Of course, if it is assumed that curvature can be represented by a particular parametric model, curvilinear associations can be addressed. For example, it might be assumed that $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.

However, experience with smoothers (e.g., Hastie & Tibshirani, 1990) suggest that it is not always evident which parametric model provides a good reflection of the data.

Using a special case of a wild bootstrap method derived by Stute, Manteiga and Quindimil (1998), it is possible to test

$$H_0 : m(X) = \theta, \qquad (1)$$

where θ is some unknown measure of location. That is, the hypothesis is that the regression line is a straight, horizontal line having intercept θ. Wilcox (2001) reported simulation results indicating that good control over the probability of a Type I error is achieved when θ is taken to be the population mean of the Y values, μ_y. A seemingly natural way of robustifying this method is to replace the mean with some robust estimator, and a 20% trimmed mean was suggested by (Wilcox, 2003). One goal here is to describe situations where using a 20% trimmed mean, control over the probability of a Type I error is very poor, but when using the usual mean, satisfactory control is maintained. (Using the median of the Y values can also result in poor control over the probability of a Type I error.)

Wilcox (2006) suggested a flexible method for testing

$$H_0 : \tau \equiv 1, \qquad (2)$$

the hypothesis that the error term is homoscedastic. The success of the wild bootstrap test of (1) suggests an alternative approach to testing (2). The second goal in this paper is to report simulation results on this alternative approach and to describe situations where it has more power than the approach studied by Wilcox (2006).

Description of the Methods
Testing (1)

The wild bootstrap test of (1) is applied as follows. Let $(X_1, Y_1), ..., (X_n, Y_n)$ be a random sample of n points from some unknown bivariate distribution. Let $\overline{Y}$ be the usual sample mean based on the Y values. Fix j and set $I_i = 1$ if $X_i \le X_j$, otherwise $I_i = 0$, and let

$$R_j = \frac{1}{\sqrt{n}} \sum I_i r_i, \qquad (3)$$

where $r_i = Y_i - \overline{Y}$. The test statistic is the maximum absolute value of all the R_j values:

$$D = \max |R_j|, \qquad (4)$$

where the maximum is over all j. The critical value is computed as follows. Generate n observations from a uniform distribution and label the results $U_1, ..., U_n$. For $i = 1, ..., n$, set $V_i = \sqrt{12}(U_i - .5)$, $r_i^* = r_i V_i$, and $Y_i^* = \overline{Y} + r_i^*$. Then based on the n pairs of points $(X_1, Y_1^*), ..., (X_n, Y_n^*)$, compute the test statistic as described in the previous paragraph and label it D^*. Repeat this process B times and label the resulting (bootstrap) test statistics $D_1^*, ..., D_B^*$. Finally, put these B values in ascending order, which are labeled $D_{(1)}^* \le ... \le D_{(B)}^*$. Then, the α level critical value is $D_{(u)}^*$, where $u = (1 - \alpha)B$ rounded to the nearest integer. That is, reject if $D \ge D_{(u)}^*$.

The corresponding p-value is the proportion of D_b^* values, among the B bootstrap samples, for which $D < D_b^*$.

Let $Y_{(1)} \leq \ldots \leq Y_{(n)}$ be the Y_i values written in ascending order and let g be equal to .2n rounded to the nearest integer. Then the 20% trimmed mean of the Y values is

$$\bar{Y}_t = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} Y_{(i)} .$$

An alternative test of (1) is obtained simply by replacing the usual sample mean, $\bar{Y}$, with $\bar{Y}_t$.

Testing (2).

Now a test of (2) is described that is based on a simple modification of the method in Wilcox (2006). Let $m(X)$ be the conditional median of Y, given X. The first step is to approximate $m(X)$ using what is called a running interval smoother, which is applied follows. Let f, called a span, be some constant to be chosen and let M be the median of the values $X_1, \ldots, X_n$. The median absolute deviation (MAD) measure of dispersion is the median of the values $|X_1 - M|, \ldots, |X_n - M|$. The point X is said to be close to X_i if

$$|X_i - X| \leq f \frac{MAD}{.6745} .$$

Under normality, MAD/.6745 estimates the standard deviation, in which case X is close to X_i if X is within f standard deviations of X_i. Let

$$N(X_i) = \{j : |X_j - X_i| \leq f \times MADN\} ,$$

where for convenience, MADN is MAD/.6745. That is, N(X_i) indexes the set of all X_j values that are close to X_i. Then $m(X_i)$ is taken to be the median of the Y_j values such that

$j \in N(X_i)$. Generally, a good choice for the span is $f = .8$ (Wilcox, 2005), and this value is used here exclusively. Let $v_i = |Y_i - m(X_i)|$ ($i = 1, \ldots, n$) be the absolute residuals. When (2) is true (there is homoscedasticity), the regression line between X and v should be a straight, horizontal line, which can be tested with the method.

Wilcox (2006) suggested two alternative methods for testing (2). Let β be the slope of the regression line between the X_i and v_i values. Then, a test of (2) corresponds to testing

$$H_0 : \beta = 0 . \qquad (5)$$

Alternatively, if ρ is some correlation between X_i and v_i, then $H_0 : \rho = 0$ should be true. A natural strategy is to use least squares regression or Pearson's correlation, but this was found to be unsatisfactory. What was found to perform well in simulations was a test of (5) using the Theil (1950) and Sen (1968) regression estimator in conjunction with a percentile bootstrap method, or an approach based on a so-called Winsorized correlation coefficient.

Consider first the regression method. The goal is to test the hypothesis that the (population) regression line between v and X is horizontal. To elaborate on the Theil-Sen estimator, for any $i < i'$, for which $X_i \neq X_{i'}$, let

$$S_{ii'} = \frac{v_i - v_{i'}}{X_i - X_{i'}} .$$

The Theil-Sen estimate of the slope is $\hat{\beta}_{ts}$, the median of all the slopes represented by $S_{ii'}$. Let β_{ts} be the population slope estimated by $\hat{\beta}_{ts}$. To test H_0: $\beta_{ts} = 0$, it currently seems that a basic percentile bootstrap method performs relatively well. In particular, a bootstrap sample is obtained by randomly sampling, with replacement, n pairs of points from $(v_1, X_1), \ldots, (v_n X_n)$. Let $\hat{\beta}^*$ be the Theil-Sen

estimate of β_{ts} based on this bootstrap sample. Repeat this bootstrap process B times yielding $\hat{\beta}_1^*,...,\hat{\beta}_B^*$. Let $\hat{\beta}_{(1)}^* \leq ... \leq \hat{\beta}_{(B)}^*$ be the bootstrap estimates written in ascending order. Let $l = \alpha/2$, rounded to the nearest integer, and $u = B - l$. Then,

$$(\hat{\beta}_{(l+1)}^*, \hat{\beta}_{(u)}^*)$$

is an approximate $1 - \alpha$ confidence interval for β_{ts}. Let $\hat{p}$ be the proportion of bootstrap estimates less than zero. Then, a p-value is $2\min(\hat{p}, 1 - \hat{p})$.

As for the Winsorized correlation approach, set $Y_{i1} = v_i$ and $Y_{i2} = X_i$ ($i = 1,...,n$). Next, Winsorize the Y values. That is, for fixed j, let $Y_{(1)j} \leq ... \leq Y_{(n)j}$ be the n values written in ascending order, and let

$$W_{ij} = Y_{(g+1)j} \text{ if } Y_{ij} \leq Y_{(g+1)j}$$
$$W_{ij} = Y_{ij} \quad \text{if } Y_{(g+1)j} < Y_{ij} < Y_{(n-g)j}$$
$$W_{ij} = Y_{(n-g)j} \text{ if } Y_{ij} \leq Y_{(n-g)j}$$

where g=[γn], γ ($0 \leq \gamma < .5$) is the amount of Winsorizing to be done and [.] is the greatest integer function. Here, $\gamma = .2$ is used. Then the estimate of ρ_w, the sample Winsorized correlation between R and X, is just Pearson's correlation based on the Winsorized values. That is, estimate ρ_w with

$$r_w = \frac{\sum (W_{i1} - \bar{W}_1)(W_{i2} - \bar{W}_2)}{\sqrt{\sum (W_{i1} - \bar{W}_1)^2 \sum (W_{i2} - \bar{W})^2}}$$

To test $H_0 : \rho_w = 0$, compute

$$T_w = r_w \sqrt{\frac{n-2}{1-r_w^2}},$$

and reject if $|T_w| \geq t_{1-\alpha/2}$, the $1 - \alpha/2$ quantile of Student's t distribution with $\nu = h - 2$ degrees of freedom, where h=n-2g.

A Simulation Study

Wilcox (2006) studied the small-sample properties of the method using simulations where both the X and Y values were generated from one of four g-and-h distributions (Hoaglin, 1985), one of which was normal. If Z has a standard normal distribution, then

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2/2) \quad \text{if } g > 0$$

$$W = \exp(hZ^2/2) \text{ if } g = 0$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used 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 considered. 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	K_1	K_2
0.00	0.00	0.00	3.00
0.00	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Although skewed distributions were considered, it turns out that when Y has other skewed distributions, not considered by Wilcox (2001), control over the probability of a Type I error is poor when using a 20% trimmed mean, but control remains good when using the mean instead. Suppose, for example, Y has a chi-squared distribution with 1 degree of freedom. Then, when X has a standard normal distribution and n=30, the actual Type I error probability is approximately .54 when testing at the .05 level (based on simulations with 1,000 replications). In contrast, when using the mean of Y, the Type I error probability is approximately .034. Problems remain with 3 degrees of freedom but they become negligible when the degrees of freedom are increased to 5. If Y has a lognormal distribution, the Type I error probability is .25 and .035 using a 20% trimmed mean and mean, respectively. So, it is evident that when Y has a sufficiently skewed distribution, using a 20% trimmed mean can be disastrous.

Now consider the problem of testing (2) with the wild bootstrap method   First consider the exact same conditions considered by Wilcox (2001), where observations were generated with either $m(X) = X$ or $m(X) = X^2$.   Table 2 shows the estimated probability of a Type I error when testing at the .05 level with n=30. Again the estimates are based on 1,000 replications with B=500. (From Robey and 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 one wants power to be .9 when testing at the .05 level and the true α value differs from .05 by .025, then 976 replications are required.)

Table 2. Type I Error Rates, n=30, α =.05

X		ε			
g	h	g	h	$m(X)=X$	$m(X)=X^2$
0.0	0.0	0.0	0.0	.036	.038
0.0	0.0	0.0	0.2	.040	.032
0.0	0.0	0.2	0.0	.046	.028
0.0	0.0	0.2	0.2	.041	.028
0.0	0.2	0.0	0.0	.039	.042
0.0	0.2	0.0	0.2	.039	.034
0.0	0.2	0.2	0.0	.043	.038
0.0	0.2	0.2	0.2	.043	.032
0.2	0.0	0.0	0.0	.040	.041
0.2	0.0	0.0	0.2	.037	.025
0.2	0.0	0.2	0.0	.044	.038
0.2	0.0	0.2	0.2	.033	.027
0.2	0.2	0.0	0.0	.034	.039
0.2	0.2	0.0	0.2	.033	.035
0.2	0.2	0.2	0.0	.043	.036
0.2	0.2	0.2	0.2	.040	.034

As indicated in Table 2, among all situations considered, the estimated probability of a Type I error ranged between .027 and .044. From Wilcox (2006), when using the Theil-Sen estimator, the estimates ranged between .030 and .067. As for the method based on the Winsorized correlation, the estimates ranged between .021 and .050. Because generating observations from a chi-squared distribution with 1 degree of freedom, or a lognormal distribution, caused problems when using the wild bootstrap method with a 20% trimmed mean, these two distributions were also

considered when using the wild bootstrap to test (2). The estimated probability of a Type I error for these two cases were .037 and .036, respectively.

A practical issue is how the power of the method compares to the power of the methods studied by Wilcox (2006). Checks revealed that the method can have more or less power depending on the nature of the heteroscedasticity. For example, if $Y = X^2 + |.2X|\varepsilon$, with both X and ε having standard normal distributions, the wild bootstrap method has power .34, versus .09 and .04 when using the Winsorized correlation or the Theil-Sen estimator, respectively. But if $Y = X^2 + |X+1|\varepsilon$, the estimated power for these three methods is .39, .51 and .35. Currently, it is unclear how best to characterize the situations where the wild bootstrap method will have more or less power. All that can be said is that given some data, the choice of method can make a practical difference.

Conclusion

For a wide range of situations, inferences based on a 20% trimmed mean, rather than a mean, can have considerable practical value in terms of both Type I errors and power (Wilcox, 2003, 2005). But, it is evident that when testing (1) with a wild bootstrap, using a 20% trimmed mean can be disastrous. Perhaps there is some modification of the wild bootstrap that both corrects this problem and has some practical advantage over using means, but this remains to be seen.

References

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

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring data tables, trends, and shapes*. New York: Wiley.

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.

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

Stute, W., Manteiga, W. G. & Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association, 93*, 141-149.

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

Wilcox, R. R. (2001). Detecting nonlinear associations plus comments on testing hypotheses about the correlation coefficient. *Journal of Educational and Behavioral Statistics, 26*, 73-84.

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.

Wilcox, R. R. (2006). Testing the hypothesis of a homoscedastic error term in simple, nonparametric regression. *Educational and Psychological Measurement, 66*, 85-92.