METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

Robust Correlation Coefficients That Deal With Bad Leverage Points

Rand R. Wilcox [1]

[1] *Department of Psychology, University of Southern California, Los Angeles, CA, USA.*

Methodology, 2023, Vol. 19(4), 348–364, https://doi.org/10.5964/meth.11045

Received: 2023-01-11 • **Accepted:** 2023-11-22 • **Published (VoR):** 2023-12-22

Handling Editor: Katrijn Van Deun, Tilburg University, Tilburg, The Netherlands

Corresponding Author: Rand R. Wilcox, Department of Psychology, University of Southern California, 3620 S. McClintock Ave Los Angeles, CA 90089-1061, USA. E-mail: rwilcox@usc.edu

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Consider the usual linear regression model. A well-known concern is that a bad leverage point, which is a type of outlier, can result in a poor fit to the bulk of the data, even when using any one of many robust regression estimators. In terms of measuring the strength of the association, bad leverage points can mask a strong association among the bulk of the data, and bad leverage points can suggest a strong association when in fact there is, in general, a weak association. This issue can be addressed by using an analog of Pearson's correlation that is eliminates outliers. But this approach can have a negative impact because it eliminates what are known as good leverage points. The paper suggests a class of robust measures of association that deals with this issue.

Keywords

robust methods, measures of association, outliers, leverage points

A well-established result is that outliers can wreak havoc on Pearson's correlation resulting in a poor and misleading understanding of the nature of the association among the bulk of the data (e.g., Kim et al., 2015; Niven & Deutsch, 2012). These results are related to the fact that the usual estimate of Pearson's correlation has a breakdown point of only $1/n$, where n is the sample size and the breakdown point is the minimum proportion of points that must be altered to make the estimate arbitrarily large or small (e.g., Wilcox, 2022). Moreover, the population Pearson correlation, ρ, has an unbounded influence

function. This means that even a small departure from a bivariate normal distribution can alter ρ in a manner that masks a strong association as illustrated in Wilcox (2022).

Numerous estimators have been derived that are aimed at dealing with outliers or extreme values when measuring the strength of an association. Extensive comparisons of 11 such estimators are reported by Li (2022). Several of these estimators are effective at dealing with the limitations of Pearson's correlation. But all of them have a common feature that can be a practical concern: they do not make a distinction between good and bad leverage points.

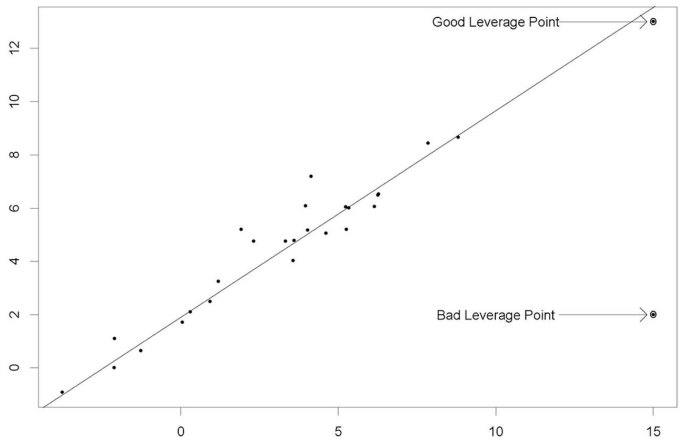
Consider the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and assume that for the bulk of these points

$$Y = \beta_0 + \beta_1 X + \varepsilon \tag{1}$$

where the slope and intercept are unknown and ε is a random variable having some unknown distribution. The point (X_i, Y_i) is called a leverage point if X_i is an outlier among the values $X_1, \ldots, X_n$. Let b_0 and b_1 be estimates of β_0 and β_1, respectively and let $r_i = Y_i - b_0 - b_1 X_i$ $(i = 1, \ldots, n)$ denote the residuals. If r_i is an outlier among $r_1, \ldots, r_n$ and simultaneously (X_i, Y_i) is a leverage point, the point (X_i, Y_i) is a bad leverage point. If (X_i, Y_i) is a leverage point, but r_i is not an outlier, the point (X_i, Y_i) is a good leverage point. Roughly, a good leverage point is a point that is reasonably consistent with the linear model for the bulk of the data given by (1) as illustrated in Figure 1. Bad leverage points can negatively impact Pearson's correlation, Kendall's tau and Spearman's rho as illustrated in the Concluding Remarks.

Figure 1

An Illustration of a Good and Bad Leverage Point



Let b_1 denote the least squares estimator for the slope and consider the squared standard error of b_1:

$$\frac{\sigma^2}{\sum\left(X_i - \bar{X}\right)^2}, \tag{2}$$

where σ^2 is the variance of the error term in (1). Note that a leverage point inflates the denominator, which in turn lowers the standard error. However, even a single bad leverage point can result in a poor fit to bulk of the points.

Many robust regression estimators have been derived that have a reasonably high breakdown point. However, all of these estimators, including estimators with the highest possible breakdown point, .5, can be unduly impacted by bad leverage points (Wilcox, 2022). What is needed is a method that retains good leverage points and eliminates bad leverage points.

There are two broad classes of robust measures of association. The first are measures that deal with outliers among the marginal distributions and the second deals with outliers in a manner that takes into account the overall structure of the data cloud (Wilcox, 2022). But extant versions do not deal with bad leverage points. This is because these measures are not directly tied to any particular regression estimator. Here, the approach is to first assume that (1) is a reasonable model of the association among the bulk of the participants. Next, use a method for estimating the slope and intercept in a manner that avoids the deleterious impact of bad leverage points and then use this fit to compute an analog of Pearson's correlation.

The paper is organized as follows. The remainder of this section reviews the method used to detect bad leverage points. The second section, Proposed Measures of Association, reviews a method for detecting bad leverage points that is used here. Section 3, Simulation Results reports on how well the inferential methods in the second section, Proposed Measures of Association, perform. The fourth section, Some Illustrations, reports simulation results on how well the inferential methods in the section, Simulation Results, perform. In contrast to the results reported by Li (2022), as well as Yuan and Mackinnon (2014), the simulation results reported here include situations where the distributions are skewed, which will be seen to be an important issue.

A major advance toward the goal of detecting bad leverage points was derived by Rousseeuw and van Zomeren (1990). Their method begins by fitting a regression line using the least median of squares (LMS) estimator. That is, the slope and intercept are estimated by the values b_1 and b_0, respectively, that minimize the median of the squared residuals. This estimator has the highest possible breakdown point, .5, where the breakdown point refers to the minimum number of points that must be altered to make the estimates arbitrarily large or small. Essentially, the breakdown point reflects the sensitivity of an estimator to outliers. A natural conclusion at the time was that

because the LMS estimator has the highest possible breakdown point, it will provide a reasonable estimate of the regression line given by (1) even when there are leverage points. However, this is not necessarily the case.

Many robust regression estimators have been derived that have a reasonably high breakdown point. Nevertheless, generally these estimators can be substantially influenced by a few bad leverage points (Wilcox, 2022). That is, a few bad leverage points cannot make the estimates arbitrarily large, but they can alter the estimate of slope and intercept to the point that a poor fit to the bulk of the data is obtained. One consequence is that the method derived by Rousseeuw and van Zomeren can miss bad leverage points.

Wilcox and Xu (2023) suggested a slight modification of the Rousseeuw and van Zomeren method that deals with the issue just described. The method begins by removing all leverage points and estimating the slope and intercept using some robust regression estimator. No single estimator dominates in terms of efficiency, but two that stand out are the MM-estimator derived by Yohai (1987) and the estimator derived by Theil (1950) and Sen (1968). The MM-estimator has the highest possible breakdown point, .5, while the Theil–Sen estimator has a breakdown point of .29. Both of these estimators are more efficient than the LMS estimator. A possible practical concern with the MM-estimator is that situations are encountered where the iterative estimation method fails. The Theil–Sen estimator avoids this problem and is used here with the understanding that there might be situations where the MM-estimator, or even some other robust estimator, offers a practical advantage.

The Theil–Sen estimator is computed as follows. Let

$$S_{ij} = \frac{Y_i - Y_j}{X_i - X_j}, \tag{3}$$

for every $i < j$. The estimate of the slope, b_1, is the median of the S_{ij} values. The intercept is estimated with $b_0 = M_y - b_1 M_x$, where M_y and M_x are the medians based on $(Y_1, \ldots, Y_n)$ and $(X_1, \ldots, X_n)$, respectively.

Note that by deleting all leverage points, the deleterious impact of bad leverage points has been eliminated in which case values for b_0 and b_1, based on a robust regression estimator, provide an estimate of the parameters in (1). Next, based on this fit, compute the residuals using all of the data yielding $u_1, \ldots, u_n$. If u_i is an outlier among $u_1, \ldots, u_n$, and if (X_i, Y_i) is a leverage point, decide that (X_i, Y_i) is a bad leverage point.

Here, the MAD-median rule is used to detect outliers, which is a special case of the outlier detecting method in Rousseeuw and van Zomeren (1990). To elaborate, let MAD denote the median of $|X_1 - M_x|, \ldots, |X_n - M_x|$. Then X_i is declared an outlier if

$$\frac{|X_i - M_x|}{MAD/.6745} \geq 2.24. \tag{4}$$

Under normality, MAD/.6745 estimates the population standard deviation.

The Proposed Measures of Association

Let $\tau^2(\widehat{Y})$ and $\tau^2(Y)$ denote some measure of variation associated with $\widehat{Y}$ and Y, respectively, where $\widehat{Y} = \beta_0 + \beta_1 X$. Explanatory power is

$$\xi^2 = \frac{\tau^2(\widehat{Y})}{\tau^2(Y)}. \tag{5}$$

From basic principles, when using the least squares regression estimator and τ^2 is taken to be the variance, ξ^2 reduces to ρ^2, where ρ is Pearson's correlation.

A robust version of ξ^2 is obtained simply by using a robust regression estimator in conjunction with some choice for τ^2 that is robust as well. There are many robust measures of variation, comparisons of which are reported by Lax (1985). One that was found to perform relatively well is the percentage bend measure of variation. The version used here has a breakdown point of .2, which is reasonably high.

The percentage bend measure of variation is computed as follows. Compute $.8n + .5$, round this value down to the nearest integer and label the result k. Let $W_i = |X_i - M_x|$, $i = 1, \ldots, n$. Let $W_{(1)} \leq \ldots \leq W_{(n)}$ be the W_i values written in ascending order. Let

$$\widehat{\omega}_\beta = W_{(k)},$$

$$Y_i = \frac{X_i - M_x}{\widehat{\omega}_\beta}$$

and let $a_i = 1$ if $|Y_i| \geq 1$. If $|Y_i| > 1$, $a_i = 0$. The percentage bend midvariance is

$$\widehat{\zeta}^2_{pb} = \frac{n\widehat{\omega}^2_\beta \sum \{\Psi(Y_i)\}^2}{\left(\sum a_i\right)^2}, \tag{6}$$

where

$$\Psi(x) = max[-1, min(1, x)].$$

Under normality, this estimator provides a very close estimate of the population variance.

In summary, the proposed robust analog of the coefficient of determination that deals with bad leverage points is computed as follows. Compute b_0 and b_1, estimates of the slope and intercept, respectively, using a robust regression estimator with bad leverage points removed. This leaves $m \leq n$ points which for notational convenience are denoted

by $(X_1, Y_1), \ldots, (X_m, Y_m)$. Let $\hat{Y}_i = b_0 + b_1 X_i$, $(i = 1, \ldots, m)$. Let $\hat{\tau}^2(\hat{Y})$ denote the percentage bend measure of variation based on $\hat{Y}_1, \ldots, \hat{Y}_m$ and let $\hat{\tau}^2(Y)$ denote the percentage bend measure of variation based on $Y_1, \ldots, Y_m$. An estimate of ξ^2 is

$$\hat{\xi}^2 = \frac{\hat{\tau}^2(\hat{Y})}{\hat{\tau}^2(Y)}. \tag{7}$$

A robust analog of ρ that deals with bad leverage points is estimated with

$$\hat{\eta} = sign(b_1)\hat{\xi} \tag{8}$$

There is a feature of $\hat{\eta}$ that should be noted. Its value can depend on which variable is taken to be the dependent variable. This is, of course, in contrast to Pearson's correlation, Spearman's rho and Kendall's tau.

Inferences About η

There is the issue of testing hypotheses and computing confidence intervals for η, the population parameter being estimated by $\hat{\eta}$. Here, the same three bootstrap methods studied by Li (2022) are considered. For two of these methods there are in fact two variations considered here that differ in how bootstrap samples are generated. In effect, five methods are considered here. The first variation is related to extant methods for making inferences about the parameters associated with a linear model. An approach that has been studied extensively is to first remove any outliers among the independent variable and then make inferences about the slope and intercept using the remaining data.

The details of the first method are as follows. Again let $(X_1, Y_1), \ldots, (X_m, Y_m)$ denote the remaining data after bad leverage points are removed. Next, generate a bootstrap sample by randomly sampling with replacement m pairs of points from $(X_1, Y_1), \ldots, (X_m, Y_m)$ yielding the bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_m^*, Y_m^*)$.

Based on a bootstrap sample, compute $\hat{\tau}^2(\hat{Y})$ and use this value in (7) and (8) yielding $\hat{\eta}^*$. Note that a bootstrap version of $\hat{\tau}^2(Y)$ is not used. If a bootstrap estimate is used, simulations indicated that control over the Type I error probability is poor.

Next, repeat the above process B times yielding $\hat{\eta}_1^*, \ldots, \hat{\eta}_B^*$. A bootstrap estimate of the squared standard error of $\hat{\eta}$ is

$$S^2 = \frac{1}{B-1}\sum(\hat{\eta}_b^* - \bar{\eta}^*)^2, \tag{9}$$

where $\bar{\eta}^* = \sum \hat{\eta}_b^*/B$. Results in Efron (1987) suggest that generally, $B = 100$ suffices and was found to perform reasonably well in simulations, but at the suggestion of a referee, $B = 1000$ is used here.

The test statistic for testing

$$H_0: \eta = 0 \tag{10}$$

is

$$W = \frac{\widehat{\eta}}{S}, \tag{11}$$

which is assumed to have a standard normal distribution when the null hypothesis is true. A $1 - \alpha$ confidence interval for η is taken to be

$$\vec{\eta} \pm z_{1-\alpha/2} S, \tag{12}$$

where $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of a standard normal distribution. Following Li, this method is labeled BSI.

The next approach is based on two variations of the basic percentile bootstrap method. The first, which is labeled method PBR, proceeds as done by method BSI, meaning that the analysis begins by removing bad leverage point and generating bootstrap samples based on the remaining data. In contrast to BSI, no estimate of the standard error is used. Another difference is that now both the numerator and denominator of (7) are computed based on a bootstrap sample. Put the bootstrap values $\widehat{\eta}_1^*, \ldots, \widehat{\eta}_B^*$ in ascending order yielding $\widehat{\eta}_{(1)}^* \leq \ldots \leq \widehat{\eta}_{(B)}^*$. Let $p^* = A/B$, where A is the number of bootstrap estimates that are less than zero. A p-value is

$$2\min\left(p^*, 1 - p^*\right) \tag{13}$$

(Liu & Singh, 1997). A $1 - \alpha$ confidence interval is

$$\left(\widehat{\eta}_{(\ell+1)}^*, \widehat{\eta}_{(u)}^*\right), \tag{14}$$

where $\ell = \alpha B/2$ rounded to the nearest integer and $u = B - \ell$.

Methods BSI and PBR reflect a common approach, in the context of a linear model, where leverage points, but not outliers among the dependent variable, are removed and a percentile bootstrap method is performed on the remaining data (Wilcox, 2022). In contrast, when working with some measure of association, bootstrap samples are based on all of the data. Otherwise, given the goal of making inferences about η, the details are exactly the same as method PBR. This alternative approach is labeled PBC.

The final two methods are based on the bias-corrected and accelerated bootstrap (BCa) interval. The main difference from the percentile bootstrap method is that BCa attempts to correct for any bias associated with an estimator as well as any skewness associated with the bootstrap distribution. This is done via two parameters. The first is based on the proportion of bootstrap estimates that are less than the observed value,

$\widehat{\eta}$. The second is aimed at adjusting the confidence interval based on how much the bootstrap distribution is skewed to the right or left. Complete computational details are in Efron (1987) and Li (2022). The practical point here is that the method is readily applied via the R function bcajack2, which can be accessed via the R package bcaboot. Again two variations are used. The first begins by removing bad leverage points (BCaR) and the second uses all of the data when generating bootstrap samples (BCaC).

Simulation Results

Simulations were used to get some sense of how well the methods in the Simulation Results section perform when dealing with sample sizes $n = 20$, 40 and 100. A few simulations were run with $n = 200$ and 300.

Data were generated based on the linear model given by (1). Three distributions were used for X and the error term, ε: standard normal, a symmetric heavy-tailed distribution and a skewed distribution with relatively high skewness and kurtosis.

An h distribution, which belongs to the family of g-and-h distributions, was used to generate data from a symmetric, heavy-tailed distribution. A value is generated from this distribution by first generating a value from a standard normal distribution yielding Z and computing

$$X = Z\exp\left(hZ^2/2\right). \tag{15}$$

The skewed distribution that was used is a lognormal distribution, which is motivated in part by various studies summarized in Wilcox (2022) looking at the degree distributions might be non-normal. This distribution is also motivated by results reported by Cain et al. (2017) who reviewed estimates of skewness and kurtosis reported in papers published in two journals: *Psychological Science* and the *American Education Research Journal*. The skewness and kurtosis of a lognormal distribution are 6.185 and 113.9, respectively. Based on 1,567 estimates collected by Cain et al. (2017), this level of skewness is realistic and relatively extreme. The skewness and kurtosis of a lognormal distribution is larger than 99% of the estimates reported by Cain and colleagues, suggesting that if a method performs reasonably well for this seemingly large departure from a normal distribution, this offers some assurance that it will work well in practice.

Table 1 reports $\widehat{\alpha}$, the estimated probability of a Type I error when using BSI, $\beta_0 = \beta_1 = 0$ and when testing at the .05 level. The estimates are based on 3000 replications. Bradley (1978) suggested that as a general guide, when testing at the .05 level, the actual level should be between .025 and .075. As can be seen, all of the estimates in Table 1 indicate that this criterion is satisfied. Moreover, based on the Agresti and Coull (1998) method, the .95 confidence interval for $\widehat{\alpha}$ does not contain .075 if $\widehat{\alpha} \leq .63$ and it does not contain .025 if $\widehat{\alpha} \geq .31$. Note that the largest estimate in Table 1 is .068, which occurred

when X has a lognormal distribution, ε has a standard normal distribution and $n = 100$. Increasing the sample size to $n = 200$, the estimate is .074. For $n = 300$, the estimate is .077 suggesting that BSI might not be asymptotically correct.

Table 1

Estimated Probability of a Type I Error ($\hat{\alpha}$) Using Method BSI

X	ϵ	$n = 20$ ($B = 100$)	$n = 20$	$n = 40$	$n = 100$
N	N	.062	.054	.057	.047
N	LN	.041	.036	.023	.037
N	H	.048	.051	.043	.051
H	N	.065	.053	.054	.058
H	LN	.047	.048	.029	.029
H	H	.057	.034	.050	.047
LN	N	.063	.049	.051	.068
LN	LN	.037	.040	.023	.034
LN	H	.047	.036	.043	.065

Note. Testing at the .05 level. $B = 1000$ except where noted. $U = \varepsilon$. N = normal, LN = lognormal, H = h distribution.

As for method PBR, situations were found where it performed in a reasonably accurate manner, but situations were found where it performed poorly. For example, when the independent variable has a standard normal distribution, the estimated Type I error probabilities ranged between .051 and .058 for the three sample sizes and the three distributions used for ε. However, when the independent variable has a lognormal distribution, it performed poorly. When ε has a normal distribution, the estimates for the three sample sizes are .072, .092 and .100, respectively. When ε has a lognormal distribution as well, the estimates are .072, .061 and .095. That is, method PB deteriorates as the sample size increases. Consequently, this method is not considered further.

Table 2 reports the results for method PBC. A limitation of method PBC is that when $n = 20$, this often resulted in numerical errors associated with estimates based on a bootstrap sample. Consequently, results for $n = 20$ are not reported. In contrast to method PBR, the estimates of α are close to the nominal level for all of the situations considered. Moreover, for the situations where PBR fails as the sample size increases, this was not the case using PBC. For example, with $n = 200$ and where both X and ε have lognormal distributions, the estimate is .045 using PBC. It was previously noted that when X has a lognormal distribution and ε has a normal distribution, the estimated probability of a Type I error using BSI is .074 when $n = 200$ and .077. Using PBC, the estimate is .045.

Table 2

Estimated Probability of a Type I Error ($\hat{\alpha}$) Using Method PBC

X	ϵ	$n = 40$	$n = 100$
N	N	.041	.047
N	LN	.041	.042
N	H	.041	.045
H	N	.038	.044
H	LN	.038	.039
H	H	.036	.041
LN	N	.038	.042
LN	LN	.043	.055
LN	H	.037	.039

Note. Testing at the .05 level. $B = 1000$. N = normal, LN = lognormal, H = h distribution.

Table 3 reports results when using BCaR and BCaC. A positive feature of BCaR is that it improves on method PBR, but when the independent variable is lognormal, it performed poorly. Indeed, like method PBR, as the sample size increases, its ability to control the Type I error probability deteriorates. Notice that when X has a lognormal distribution and ε has the h distribution, the estimate is .107 when $n = 100$. Increasing the sample size to $n = 200$, the estimate is .117.

Table 3

Estimated Probability of a Type I Error ($\hat{\alpha}$) Using Methods BCaR and BCaC

		BCaR			BCaC	
X	ϵ	$n = 20$	$n = 40$	$n = 100$	$n = 40$	$n = 100$
N	N	.057	.051	.048	.040	.042
N	LN	.065	.061	.059	.050	.047
N	H	.051	.054	.052	.041	.053
H	N	.055	.064	.065	.032	.046
H	LN	.054	.061	.065	.042	.051
H	H	.053	.056	.067	.043	.040
LN	N	.060	.082	.080	.027	.045
LN	LN	.062	.074	.085	.045	.064
LN	H	.077	.086	.107	.030	.046

Note. Testing at the .05 level. $B = 1000$. N = normal, LN = lognormal, H = h distribution.

Both PBC and BCaC were found to control the probability of a Type I error reasonably well for all situations considered. BSI performed reasonably well when $n \leq 100$ and has the advantage of avoiding computational issues associated with PBC and BCaC when $n = 20$. But the extent these three methods deal with bad leverage points is unclear. The h distribution and the lognormal distribution have a tendency to generate outliers. That is, situations are being considered where leverage points are highly likely to occur as well as regression outliers, meaning outliers among the residuals. But a criticism is that the likelihood of a bad leverage point is relatively low. To deal with this, the simulations were repeated only now two leverage points were added to the data, namely (4, 4) and (5, 5). The lognormal distribution was shifted to have a median equal to zero so that the value 4 is an outlier. The resulting estimates of the Type I error probability are reported in Table 4. As can be seen, BCaC always performed about as well or better than PBC. BSI is a bit better than BCaC in some situations but worse in others. If the goal is have a Type I error probability reasonably close but less than the nominal level, BCaC is best.

Table 4

Estimated Probability of a Type I Error ($\hat{\alpha}$) When There Are Two Bad Leverage Points Included Using Methods BSI, PBC and BCaC

X	ϵ	BSI (n = 20)	BCaC (n = 40)	PBC (n = 40)
N	N	.058	.039	.041
N	LN	.040	.045	.039
N	H	.065	.044	.038
H	N	.062	.035	.035
H	LN	.039	.043	.037
H	H	.062	.045	.035
LN	N	.063	.033	.030
LN	LN	.037	.034	.033
LN	H	.065	.034	.028

Note. Testing at the .05 level. $B = 1000$. N = normal, LN = lognormal, H = h distribution.

The next set of simulations focused on the ability to get a reasonably accurate $1 - \alpha = .95$ confidence interval for η when $\eta \neq 0$. Now $\beta_0 = 0$ and $\beta_1 = 1$ are used. The actual value of η was determined by generating a random sample of size 100, computing $\widehat{\eta}$ and repeating this process 1000 times. The mean of the resulting 1000 estimates is taken to be the true value of η. Another approach is to use one large sample size, say 10000. The Theil–Sen estimator is easily computed when $n = 1000$, but for $n = 10000$ this is no longer the case. In practice, when n is extremely large, some alternative robust estimator can be required. A good choice is the MM-estimator derived by Yohai (1987).

Table 5 reports the results for methods BSI and BCaR and Table 6 reports results using methods PBC and BCaC. Method BCaR was included because in contrast to the results in Table 3, it performed reasonably well even when X has a lognormal distribution. Indeed, in a variety of situations, it performed better than method BSI. The main concern with BSI is that is that the estimates are less than .025 in some situations and the estimates decrease as n gets large, another indication that BSI is not asymptotically correct in some situations.

Table 5

Estimates of α When Computing a 1 − α = .95 Confidence Interval Using Methods BSI and BCaR

		$n = 20$		$n = 40$		$n = 100$		
X	**ϵ**	**BSI**	**BCaR**	**BSI**	**BCaR**	**BSI**	**BCaR**	**η**
N	N	.024	.054	.027	.054	.047	.068	.698
N	LN	.038	.050	.022	.051	.021	.051	.645
N	H	.032	.052	.019	.064	.010	.059	.636
H	N	.024	.058	.012	.070	.047	.065	.720
H	LN	.035	.052	.025	.056	.023	.056	.637
H	H	.028	.060	.015	.066	.015	.056	.643
LN	N	.031	.061	.025	.063	.033	.068	.597
LN	LN	.035	.057	.027	.052	.030	.067	.513
LN	H	.032	.051	.025	.060	.023	.066	.520

Note. η differs from zero. N = normal, LN = lognormal, H = h distribution.

Table 6

Estimates of α When Computing a 1 − α = .95 Confidence Interval Using Methods PBC and BCaC

		PBC		BcaC		
X	**ϵ**	$n = 40$	$n = 100$	$n = 40$	$n = 100$	**η**
N	N	.012	.025	.055	.054	.698
N	LN	.017	.031	.048	.049	.645
N	H	.021	.034	.052	.047	.636
H	N	.011	.022	.056	.062	.597
H	LN	.013	.023	.053	.050	.513
H	H	.014	.023	.059	.056	.520
LN	N	.011	.018	.047	.067	.720
LN	LN	.010	.013	.044	.057	.637
LN	H	.012	.020	.046	.054	.643

Note. η differs from zero. N = normal, LN = lognormal, H = h distribution.

As indicated in Table 6, the estimates of α, when using PBC, are well below the nominal level when $n = 40$. For $n = 100$, the estimates are closer to the nominal level, but the method improves rather slowly as n increases. All indications are that BCaC performs better than PBC.

Some Illustrations

Of course, bad leverage points are not always an issue. But the reality is that situations are encountered where bad leverage points are a serious concern as illustrated here.

The first example is based on data dealing with measures of reading ability. The sample size is $n = 81$. The goal is to understand the association between a measure of speeded naming for letters and a measure of speeded naming for digits. Figure 2 shows a scatterplot of the data. Points marked with o are bad leverage points and points marked with * are good leverage points. The scatterplot clearly suggests that for the bulk of the points there is a positive association. Pearson's correlation, using all of the data, is .106. Testing the hypothesis that Pearson's correlation is zero, the p-value is .01. (The bootstrap-t method in Wilcox, 2022, Section 9.1 was used, which compares well to the methods studied by Bishara & Hittner, 2012.) Using Spearman's rho and Kendall's tau, the estimates are .452 and .347, respectively, both of which reject at the .05 level. In contrast, $\widehat{\eta} = .768$. That is, all four measures reject at the .05 level, but $\widehat{\eta}$ paints a decidedly different picture regarding the strength of the association. The .95 confidence intervals for η are (.465, 1), (.537, 1) and (.506, 1) using BSI, PBC and BCaC, respectively. Method PBC has the shortest confidence interval, which was somewhat unexpected because in the simulations, the estimate of α is smaller when using PBC compared to BCaC.

The next example is based on data reported by Rousseeuw and Leroy (1987, p. 27) that deals with the logarithm of the effective temperature at the surface of 47 stars versus the logarithm of its light intensity. A scatterplot of the data is shown in Figure 3. Pearson's correlation is $-.21$, while Spearman's correlation, Kendall's tau and $\widehat{\eta}$ are .295, .250 and .607, respectively. The corresponding p-values are .136, .132 and less than .001. The .95 confidence intervals using BSI, PBC and BCaC are (.352, .861), (.360, .844) and (.326, .795), respectively. In this case, BCaC has the shortest confidence interval.

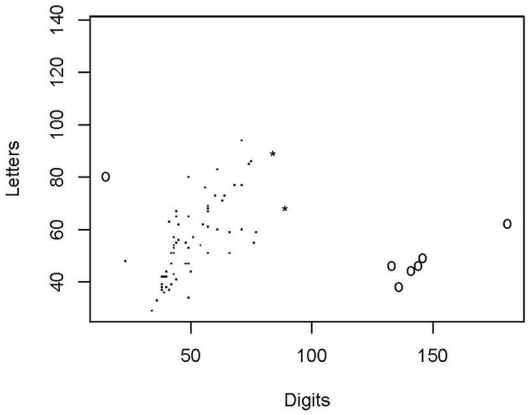
The next illustration is based on measures of cortisol levels taken upon awakening and measured again 30–45 minutes. Past studies indicate that Time 2 measures tend to be higher than the Time 1 measures. The extent this is the case has been found to be associated with various measures of stress. The goal here is to understand the strength of the association between these two measures.

Figure 4 shows a scatterplot of the data. The sample size is 101. The data stem from a study of an intervention program aimed at improving the emotional and physical wellbeing of older adults. Pearson's correlation, Spearman's rho and Kendall's tau are

.490, .494 and .358, respectively. In contrast, $\widehat{\eta} = .583$. All four measures are significant at the .01 level, but clearly the three bad leverage points have an impact on the first three correlation coefficients. The .95 confidence intervals using BSI, PBC and BCaC are (.359, .806), (.357, .757) and (.369, .763), respectively. BCaC has a slightly shorter confidence interval than PBC.

Figure 2

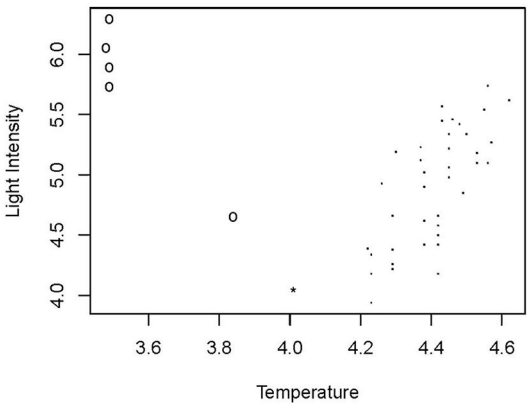
Scatterplot of the Reading Data



Note. Bad leverage points are indicated by o, good leverage points are indicated by *.

Figure 3

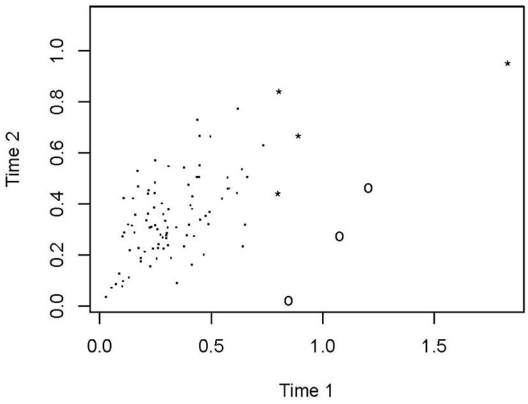
Scatterplot of the Star Data



Note. Bad leverage points are indicated by o, good leverage points are indicated by *.

Figure 4

Scatterplot of the Cortisol Data



Note. Bad leverage points are indicated by o, good leverage points are indicated by *.

Concluding Remarks

It would be convenient to have a single method that generally performs about as well or better than competing techniques. Generally, PBC and BCaC perform relatively well, but computational issues eliminate these two approaches when dealing with a small sample size ($n = 20$). For $n \geq 40$, both PBC and BCaC are good choices. BCaC performs better than PBC in general. An argument for PBC might be that a p-value is readily computed. A p-value does not indicate the probability of making a correct decision about whether η is positive or negative. But in the context of Tukey's three decision rule (e.g., Jones & Tukey, 2000), it reflects the strength of the empirical evidence that a decision can be made. BCaC can be used with any choice for α, in principle a p-value can be computed, but this is more easily done using PBC. Also, simulations suggest that BCaC will yield a shorter confidence interval than PBC, but the illustrations demonstrate that this is not necessarily the case.

There are many variations of the method used here, each based on some robust regression estimator. Presumably the use of the Theil–Sen estimator does not dominate all other robust regression estimators in terms of power, and perhaps the ability of controlling the Type I error probability, when testing (10). The only suggestion here is that the Theil–Sen estimator is reasonably good choice for general use.

The illustrations suggest that at a minimum, when dealing with a linear model, it is prudent to check whether there are any bad leverage points. Otherwise, there is a realistic chance that the nature of the association for the bulk of the points is completely missed.

Finally, R functions are available for dealing with bad leverage points. The R function outblp checks for bad leverage points. The regression estimator that is used is determined by the argument regfun, which defaults to the Theil–Sen estimator. The function corblp computes $\hat{\xi}$ and corblp.ci computes a confidence interval for ξ using method BSI. The function corblppb performs method PBC and corblp.bca.C performs method BCaC. Access to these functions can be achieved by sourcing the file Rallfun-v42, which can be downloaded from the Supplementary Materials.

Funding: The author has no funding to report.

Acknowledgments: The author has no additional (i.e., non-financial) support to report.

Competing Interests: The author has declared that no competing interests exist.

Data Availability: The data files used for this article are freely available and can be found in the Supplementary Materials.

Supplementary Materials

For this article, the materials provided are various R functions (i.e., outblp, corblp, corblp.ci, corblppb, corblp.bca, etc.), analyses, and illustrations for this study (see Wilcox, 2023).

Index of Supplementary Materials

Wilcox, R. R. (2023). *Rand R. Wilcox's quick files* [Datasets, R functions]. OSF. https://osf.io/xhe8u/

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *American Statistician, 52*(2), 119–126.

Bishara, A. J., & Hittner, J. A. (2012). Testing the significance of a correlation with nonnormal data: Comparison of Pearson, Spearman, transformation, and resampling approaches. *Psychological Methods, 17*(3), 399–417. https://doi.org/10.1037/a0028087

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

Cain, M. K., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behavior Research Methods, 49*(5), 1716–1735. https://doi.org/10.3758/s13428-016-0814-1

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*(397), 171–185. https://doi.org/10.1080/01621459.1987.10478410

Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*(4), 411–414. https://doi.org/10.1037/1082-989X.5.4.411

Kim, Y., Kim, T.-H., & Ergun, T. (2015). The instability of the Pearson correlation coefficient in the presence of coincidental outliers. *Finance Research Letters, 13*, 243–257. https://doi.org/10.1016/j.frl.2014.12.005

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association, 80*(391), 736–741. https://doi.org/10.1080/01621459.1985.10478177

Li, J. C.-H. (2022). Bootstrap confidence intervals for 11 robust correlations in the presence of outliers and leverage observations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 18*(2), 99–125. https://doi.org/10.5964/meth.8467

Liu, R. G., & Singh, K. (1997). Notions of limiting *P* values based on data depth and bootstrap. *Journal of the American Statistical Association, 92*(437), 266–277. https://doi.org/10.2307/2291471

Niven, E. B., & Deutsch, C. V. (2012). Calculating a robust correlation coefficient and quantifying its uncertainty. *Computers & Geosciences, 40*, 1–9. https://doi.org/10.1016/j.cageo.2011.06.021

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association, 85*(411), 633–639. https://doi.org/10.1080/01621459.1990.10474920

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

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*(2), 85–91.

Wilcox, R. R. (2022). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.

Wilcox, R., & Xu, L. (2023). Regression: Identifying good and bad leverage points. *International Journal of Statistics and Probability, 12*(1), 1–9. https://doi.org/10.5539/ijsp.v12n1p1

Yohai, V. J. (1987). High breakdown point and high efficiency robust estimates for regression. *Annals of Statistics, 15*(2), 642–656. https://doi.org/10.1214/aos/1176350366

Yuan, Y., & Mackinnon, D. P. (2014). Robust mediation analysis based on median regression. *Psychological Methods, 19*(1), 1–20. https://doi.org/10.1037/a0033820