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

PsychOpen GOLD

Original Article

Check for updates

Bootstrap Confidence Intervals for 11 Robust Correlations in the Presence of Outliers and Leverage Observations

Johnson Ching-Hong Li [1]

[1] *Department of Psychology, University of Manitoba, Winnipeg, MB, Canada.*

Methodology, 2022, Vol. 18(2), 99–125, https://doi.org/10.5964/meth.8467

Received: 2022-02-28 • **Accepted:** 2022-06-01 • **Published (VoR):** 2022-06-30

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

Corresponding Author: Johnson Ching-Hong Li, P508 Duff Roblin Building, Department of Psychology, University of Manitoba, Winnipeg, MB, R3T 2N2, Canada. E-mail: Johnson.Li@umanitoba.ca

Supplementary Materials: Materials [see Index of Supplementary Materials]

Abstract

Researchers often examine whether two continuous variables (X and Y) are linearly related. Pearson's correlation (r) is a widely-employed statistic for assessing bivariate linearity. However, the accuracy of r is known to decrease when data contain outliers and/or leverage observations, a circumstance common in behavioral and social sciences research. This study compares 11 robust correlations with r and evaluates the associated bootstrap confidence intervals [bootstrap standard interval (BSI), bootstrap percentile interval (BPI), and bootstrap bias-corrected-and-accelerated interval (BCaI)] across conditions with and without outliers and/or leverage observations. The simulation results showed that the median-absolute-deviation correlation (r-MAD), median-based correlation (r-MED), and trimmed correlation (r-TRIM) consistently outperformed the other estimates, including r, when data contain outliers and/or leverage observations. This study provides an easy-to-use R code for computing robust correlations and their associated confidence intervals, offers recommendations for their reporting, and discusses implications of the findings for future research.

Keywords

robust correlation, bootstrap confidence intervals, outliers, Monte Carlo simulation

Behavioral and social sciences researchers often examine whether or not two continuous variables (*X* and *Y*) are linearly related. Commonly known as linear correlation, this form

of bivariate relationship is important in model building and theory testing. Therefore, it is crucial to understand the statistical methods available to model such linear relationships, the limitations of those methods, and the conditions under which each method can be relied upon to provide accurate estimates. In this article I conduct a Monte Carlo simulation that compares the most common method to detect and estimate linear correlation, Pearson's *r*, to 11 other methods under conditions common in behavioral and social sciences but known to reduce the accuracy of *r*.

Bivariate linear relationships can be described in equation as (Cohen et al., 2003)

$$Z_y = \theta \cdot Z_x + e \tag{1}$$

where the parameter that measures the level of linearity between X and Y is θ, Z_y is the standardized variable of Y, Z_x is the standardized variable of X, and e is the residual or error variable that is generated from $e \sim N(0, \sigma_e)$, where $\sigma_e = \sqrt{1-\theta^2}$. In the existing literature there are many different statistical procedures for researchers to choose from detecting bivariate linearity (θ) between X and Y. Among them, Pearson's correlation (r) is arguably the most widely-employed procedure. In equation, r can be expressed as

$$r = \sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y}) / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

where (x_i, y_i) is a pair of x and y observations for the ith participant, $\bar{x}$ is the sample mean of all X scores, $\bar{y}$ is the sample mean of all Y scores, and n is the number of X-Y pairs in a sample.

In theory, using r as an estimator for the level of bivariate linearity is mathematically appropriate when X and Y follow a bivariate normal distribution, X and Y are linearly related, and X and Y do not contain outliers. However, in practice, behavioral and social sciences data often contain heavier tails and outliers (Albers, 2017; Micceri, 1989). This raises concerns about the adequacy of using r to estimate bivariate linearity. For example, Shevlyakov and Smirnov (2011) generated contaminated normal data, i.e., a portion of normal data[1], e.g., 10%, is manipulated to be outlier scores (Copas, 1988), with small (n = 20) and large (n = 1000) samples and found that r was inaccurate and sensitive to outliers. Other previous studies (Devlin et al., 1975; Niven & Deutsch, 2012) found that r was strongly biased by the presence of outliers. In an extreme case with n = 10, Gnanadesikan and Kettenring (1972) found that an additional outlier greater than 3 SDs above the mean could increase an observed r from almost zero to a large effect of .50. Chen (2006) comprehensively evaluated the performance of r when normality was

1) Shevlyakov and Smirnov's generated x-y points based on Tukey's (1960) gross error model, $f(x,y) = (1-e)N(x,y;0,0,1,1,\rho) + eN(x,y;0,0,k,k,\rho')$, where e = .10, k = 10, and $sign(\rho') = -sign(\rho)$. The second summand of this function would generate 10% very bad x-y points that are not only outliers (10 SDs), but they also correlate in opposite direction—a scenario that may not be common in behavioral and social data.

violated and outliers were present and found that *r* was not robust to these conditions. Kim et al. (2015) showed that *r* was severely influenced by the presence of outliers, stating that, "it is highly possible that researchers can wrongly conclude 100% of the time that the two variables are linearly related" with outliers in *X* and *Y* (p. 256).

To better understand how *r* is influenced by outliers, one can refer to the *r* algorithm in Equation (2), in which the sample mean of *X* (i.e., $\bar{x}$) and the sample mean of *Y* (i.e., $\bar{y}$) are required in the estimation. According to Kim et al. (2015), it is well known that statistical procedures (e.g., *r* in this case) that are based on sample means ($\bar{x}$ and $\bar{y}$) are sensitive to the presence of outliers. Behavioral and social sciences data often have heavier tails, i.e., outliers (Micceri, 1989; Yuan & Zhong, 2013), and hence, this raises concerns about the adequacy of *r* as the statistical procedure of choice for assessing bivariate linearity between *X* and *Y*. Indeed, there are 11 appealing robust correlations for assessing bivariate linearity between *X* and *Y* (Chen, 2006; Niven & Deutsch, 2012; Shevlyakov & Smirnov, 2011), but none of them has received the kind of attention and use as has *r* in behavioral and social sciences.

Previous simulations investigating the performance of *r* and robust alternatives have often focused on only one single type of outlier (e.g., 10% of both *X* and *Y* scores are outliers). However, outliers in bivariate data can further be conceptualized in four different ways: *x*-outlier, leverage observation, *y*-outlier, and leverage observation with opposite correlation. Together, these can be referred to as outliers and leverage observations (O-LO; Shevlyakov & Smirnov, 2011; Tukey, 1960[1]; Yuan & Zhong, 2013). There is no single Monte Carlo experiment that systematically and comprehensively compares the performance of *r* and the associated bootstrap confidence intervals (CIs) with that of the viable robust alternatives under these four different types of O-LO.

In light of this, the present study evaluates the performance of 11 robust correlations, as compared to *r*, via a Monte Carlo simulation study. In addition, the associated CIs are crucial for evaluating the sampling error surrounding the point estimate. Indeed, reporting CIs is required by many publication manuals and must necessarily be included in many journal articles (APA, 2010). Given that the bootstrap CIs have been found to be appropriate for many robust estimates (Ruscio, 2008; Ruscio & Mullen, 2012), they are also examined in the present study. The primary purpose of this study is to identify the most robust point estimate with the associated CIs so that behavioral and social science researchers can appropriately and accurately report and interpret bivariate linearity even when data contain O-LO.

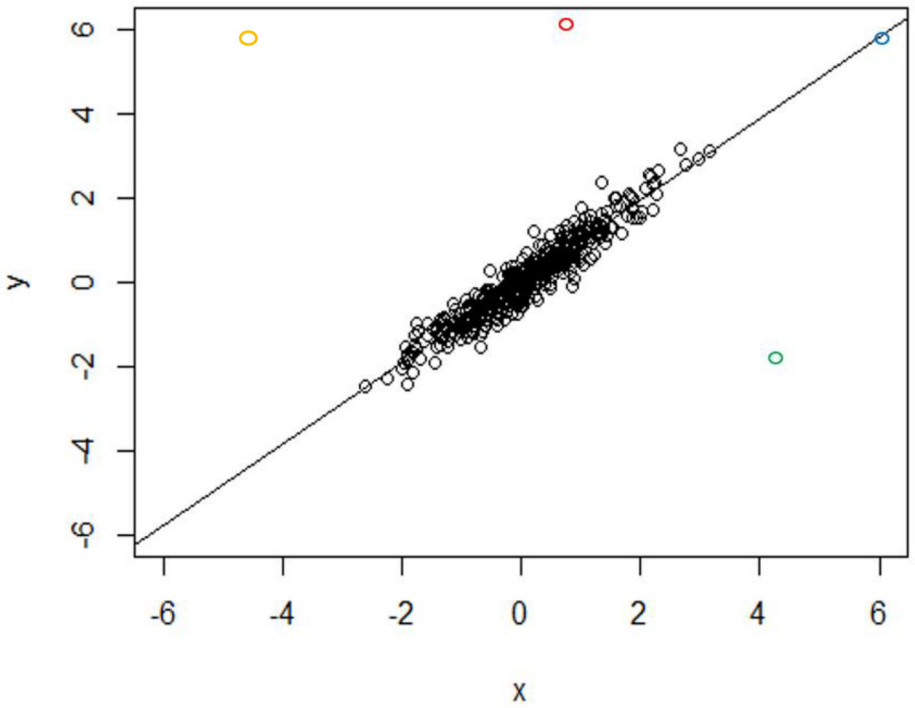
Outliers and Leverage Observations (O-LO)

According to Yuan and Zhong (2013), an outlier is defined as a point for which the *x*-coordinate (or *y*-coordinate) does not fit with the majority of the *x* (or *y*) scores; whereas a leverage observation is defined as a point where both the *x* and *y* coordinates

are outliers, but this observation still fits with the general linear pattern of all other *x*-*y* pairs. In addition to these three cases, Shevlyakov and Smirnov (2011) highlighted the difficulties presented by leverage observations with opposite correlation.

Figure 1

Scatterplot for 400 Simulated x-y Points With the Population Correlation of .95.



Note. The red point is an *y*-outlier (Case A), the blue point is a leverage observation (Case B), the green point is an *x*-outlier (Case C), and the orange point is an outlier with opposite correlation (Case D).

For Case A (outlier *y*; see Figure 1), the *y*-coordinate of point A is an outlier, and this point does not follow the general linear trend of the rest of the *x*-*y* points. This point is considered an influential point for the estimation of *r* because the sign and magnitude of the mean deviation *Y* scores, $(y_i - \bar{y})$ in the calculation of *r* in Equation (2), are substantially influenced. For Case B (leverage), given that point B is an outlier on both *x*- and *y*-coordinates but it follows the general linear trend of the remaining data points, this point is considered a leverage observation. Although this point may appear to be atypical relative to other *x*-*y* points, this point may not substantially influence the estimation of *r* because the degree to which the mean deviation *Y* scores, $(y_i - \bar{y})$, is changed is proportional to the degree to which the mean deviation *X* scores, $(x_i - \bar{x})$, is

changed in Equation (2). For Case C (outlier x), given that point C does not follow the general linear trend of the remaining data points and this point is located on the extreme end of the x-coordinate, this point is regarded as an outlier. Point C is an influential point because the degree to which $(y_i - \bar{y})$ is changed does not match in proportion with the degree to which $(x_i - \bar{x})$ is changed. Case D (leverage observation with opposite correlation) was originally proposed by Tukey (1960). Shevlyakov and Smirnov (2011) evaluated Case D in which a certain proportion (e.g., 10%) of x-y points are outliers that are themselves correlated in the direction opposite to the correlation of the remaining scores (e.g., $-\rho$, where ρ is the true correlation that relate the remaining x-y points). This condition is regarded as the most challenging condition for accurately estimating the correlation between x and y.

Four Approaches to Robust Correlations That Could Be Robust to O-LO

Approach 1: Replacement by Medians

As noted above, a first robust approach is to directly replace the linear summation of $\sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$ that is sensitive to outliers by a robust counterpart, i.e., $\sum_{i=n_l}^{n_u}[(x_i - \text{median}(x))(y_i - \text{median}(y))]$, where n_l and n_u are the lowest and highest ranked scores, respectively in a trimmed sample, e.g., 10% top and 10% bottom are trimmed as suggested in Gnanadesikan and Kettenring (1972). Next, replacing the mean deviation component, $\sum_{i=1}^{n}(x_i - \bar{x})^2\cdot\sum_{i=1}^{n}(y_i - \bar{y})^2$, by a robust counterpart $\sum_{i=n_l}^{n_u}(x_i - \text{median}(x))^2\sum_{i=n_l}^{n_u}(y_i - \text{median}(y))^2$. Consequently, a direct robust sequel to r, called a robust correlational median estimator (r_{CME}; Falk, 1998; Shevlyakov & Vilchevsky, 2002), can be expressed as

$$r_{CME} = \frac{\sum_{i=n_l}^{n_u}[(x_i - \text{median}(x))(y_i - \text{median}(y))]}{\sqrt{\sum_{i=n_l}^{n_u}(x_i - \text{median}(x))^2\sum_{i=n_l}^{n_u}(y_i - \text{median}(y))^2}} \tag{3}$$

A second type of robust correlation makes use of a deviation estimate, called the median absolute deviation, i.e., $\text{MAD}(z) = \text{median}[|z - \text{median}(z)|]$, which is more robust than the conventional standard deviation (SD) estimate. With this in mind, a robust MAD-based correlation (r_{MAD}; Pasman & Shevlyakov, 1987) can be expressed as

$$r_{MAD} = \left[\text{MAD}^2(u) - \text{MAD}^2(v)\right] / \left[\text{MAD}^2(u) + \text{MAD}^2(v)\right] \tag{4}$$

where u and v are the robust principal variables, i.e.,
$u = [x - \text{median}(x)]/\sqrt{2}\text{MAD}(x) + [y - \text{median}(y)]/\sqrt{2}\text{MAD}(y)$ and
$v = [x - \text{median}(x)]/\sqrt{2}\text{MAD}(x) - [y - \text{median}(y)]/\sqrt{2}\text{MAD}(y)$.

A third robust correlation is considered a derivative of r_{MAD}, and is known as the median correlation coefficient (r_{MED}) (Shevlyakov & Smirnov, 2011),

$$r_{MED} = \left[\text{median}^2(|u|) - \text{median}^2(|v|)\right] / \left[\text{median}^2(|u|) + \text{median}^2(|v|)\right] \quad (5)$$

where u and v are defined as in Equation (4).

A fourth robust correlation is called the biweight midcorrelation (r_{bm}; Wilcox, 2012), which first converts x_i to e_i by $e_i = \left[x_i - \text{median}(x)\right] / 9{\cdot}\text{MAD}(x)$ and y_i to f_i by $f_i = \left[y_i - \text{median}(y)\right] / 9{\cdot}\text{MAD}(y)$. Second, assign $a_i = 1$ if $-1 \le e_i \le 1$, else $a_i = 0$. Similarly, assign $b_i = 1$ if $-1 \le f_i \le 1$, else $b_i = 0$. Consequently,

$$r_{bm} = \frac{n\sum_{i=1}^{n} a_i\left[x_i - \text{median}(x)\right]\left(1 - e_i^2\right)^2 \sum_{i=1}^{n} b_i\left[y_i - \text{median}(y)\right]\left(1 - f_i^2\right)^2}{\left[\sum_{i=1}^{n} a_i\left(1 - e_i^2\right)\left(1 - 5e_i^2\right)\right]\left[\sum_{i=1}^{n} b_i\left(1 - f_i^2\right)\left(1 - 5f_i^2\right)\right]} \cdot \frac{1}{\sqrt{s_e s_f}} \quad (6)$$

where s_e and s_f are the SDs of e_i and f_i, respectively.

Approach 2: Counting Ranks or Signs of X-Y Pairs

Another approach to robust correlation involves counting the signs or ranks of x-y pairs. This approach is intended to minimize the effect of outliers and/or leverage observations on estimation of bivariate linearity. Blomqvist (1950) developed the quadrant correlation coefficient (r_Q),

$$r_Q = \sin\left(0.5{\cdot}\pi{\cdot}\sum_{i=1}^{n} \text{sign}\left[x_i - \text{median}(x)\right]{\cdot}\text{sign}\left[y_i - \text{median}(y)\right] / n\right) \quad (7)$$

where sign refers to the sign of deviations from median, and $\pi \approx 3.14156$. The component, $\sum_{i=1}^{n} \text{sign}\left[x_i - \text{median}(x)\right]{\cdot}\text{sign}\left[y_i - \text{median}(y)\right]$, is a count of the cases that when x is above (or below) the median of all xs, its paired y is also above (or below) the median of all ys. The average of the total counts (ɣ) can be scaled back to a number that measures linear association identical to the θ-metric (from -1 to +1) via Shepard's Theorem, $\sin\left(0.5{\cdot}\pi{\cdot}\gamma\right)$ (Kendall & Stuart, 1977).

A second robust correlation based on counting the x-y pairs is called Kendall's τ correlation (r_τ; Kendall, 1938),

$$r_\tau = \sin\left(0.5 * \pi * \left(n_C - n_D\right) / \left[0.5{\cdot}n(n-1)\right]\right) \quad (8)$$

where n_C refers to the number of concordant pairs for X and Y, and n_D refers to the number of discordant pairs for variables X and Y. Specifically, one can consider any x-y pairs, (x_i, y_i) and (x_j, y_j), where $i \ne j$. If the ranks for both pairs agree (i.e., either $x_i > x_j$ and $y_i > y_j$ or $x_i < x_j$ and $y_i < y_j$), then they are regarded as concordant pairs; otherwise, they are regarded as discordant pairs (i.e., either $x_i > x_j$ and $y_i < y_j$ or $x_i < x_j$ and $y_i > y_j$).

Note that Equation (8) also involves Shepard's Theorem, $\sin(0.5\cdot\pi\cdot\gamma)$, so that the value can be scaled back to the θ-metric (from -1 to +1).

In addition, Spearman's correlation (r_s; Spearman, 1904) transforms X and Y scores into rank scores, computes the distance between these rank scores, and measures bivariate linearity. That is, when there is no tie,

$$r_s = 1 - 6\sum_{i=1}^{n} d_i^2 / [n(n^2+1)] \tag{9}$$

where $d_i^2 = [\text{rank}(x_i) - \text{rank}(y_i)]^2$ is the squared difference between the rank of a score in X and a score in Y.

Approach 3: Utilizing Weighted Least Squares Estimation

Rousseeuw and Leroy (1987) proposed the weighted least squares correlation (r_{wLS}). First, a standard linear regression equation is fitted by $y_i = \beta_0 + \beta_1 x_i + e_i$. Second, $r_i = y_i - \sum_{i=1}^{n}\beta_1 x_i$ is defined as the residual score for the ith respondent. Third, $s = 1.4826[1 + 5/(n-1)]\cdot\sqrt{\text{median}(r_i^2)}$ is defined as a determining factor. Fourth, one can compute a weight for each r_i, i.e., $w_i = 1$, if $|r_i/s| \leq 2.5$, else $w_i = 0$. Consequently,

$$r_{wLS} = \sum_{i=1}^{n} w_i(x_i - \bar{x}_w)(y_i - \bar{y}_w) / \sqrt{\sum_{i=1}^{n} w_i(x_i - \bar{x}_w)^2 \sum_{i=1}^{n} w_i(y_i - \bar{y}_w)^2} \tag{10}$$

where $\bar{x}_w = \sum_{i=1}^{n} w_i x_i / \sum_{i=1}^{n} w_i$ and $\bar{y}_w = \sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$.

Approach 4: Deletion of Outliers

This approach focuses on discarding a certain percent of the top and bottom x-y pairs which are regarded as O-LO in a sample data-set. According to Huber (1981), a trimmed correlation can be presented as

$$r_{TRIM} = \left(\sum_{i=n^*+1}^{n-n^*} p_i^2 - \sum_{i=n^*+1}^{n-n^*} q_i^2\right) / \left(\sum_{i=n^*+1}^{n-n^*} p_i^2 + \sum_{i=n^*+1}^{n-n^*} q_i^2\right) \tag{11}$$

where $n^* = n\cdot(1-\pi)$ is the number of trimmed observations with π proportion, and p_i and q_i are the ith order or rank scores of variables X and Y, respectively. According to Algina et al. (2005), π is often set at 20% in order to obtain a reliable robust estimate.

Another trim-based robust correlation is called the Winsorized correlation (r_W; Wilcox, 2012). This method replaces all the values above (or below) a cut-off with a list of maximum (or minimum) scores that are left in the dataset. Specifically, researchers can rank-order the X scores such that $x(1^{st}) > x(2^{nd}) > \ldots > x(n\text{th})$. The upper and lower cut-off points become $(n\pi)$ and $(1-n\pi)$, where π is often fixed at 20% according to Wilcox (2012). Consequently, when $x(i) \leq x(n\pi)$, the Winsorized scores are converted as $w_X(i) = x(n\pi)$. When $x(n\pi) < x(i) < x(1-n\pi)$, $w_X(i) = x(i)$. When $x(i) \geq x(1-n\pi)$, $w_X(i) = x(1-n\pi)$. Repeating the same Winsorized procedure for Y, one can estimate r_w by

$$r_w = \sum_{i=1}^{n}\left[w_X(i) - \overline{w}_X\right]\cdot\left[w_Y(i) - \overline{w}_Y\right] \Big/ \sqrt{\sum_{i=1}^{n}\left[w_X(i) - \overline{w}_X\right]^2 \cdot \sum_{i=1}^{n}\left[w_Y(i) - \overline{w}_Y\right]^2} \quad (12)$$

A third trimmed correlation is known as the percentage bend correlation (r_{pbc}; Wilcox, 2012), which provides a linear estimate that is not overly sensitive to slight deviation from a parametric (normal) distribution. According to Wilcox, first compute the median deviation scores $a_i = \left|x_i - \text{median}(X)\right|$ and rank a_i in ascending order. Second, find the criterion by $\widehat{c}_x = a_{(m)}$, where $m = \lfloor (1-\pi)n \rfloor$ and $\pi = 20\%$ is defined as in (8). Third, set n_1 be the number of x_i such that $\left[x_i - \text{median}(X)\right] / \widehat{c}_x < -1$, and n_2 be the number of x_i such that $\left[x_i - \text{median}(X)\right] / \widehat{c}_x > 1$. Fourth, convert x_i to u_i by $u_i = \left(x_i - \widehat{\phi}_x\right)/\widehat{c}_x$, where $\widehat{\phi}_x = \left[\widehat{c}_x(n_2 - n_1) + b_x\right]/(n - n_2 - n_1)$ and $b_x = \sum_{i=n_1+1}^{n-n_2} x_i$. Repeat the same procedure for converting y_i to v_i through $v_i = \left(y_i - \widehat{\phi}_y\right)/\widehat{c}_y$, where $\widehat{\phi}_y = \left[\widehat{c}_y(n_2 - n_1) + b_y\right]/(n - n_2 - n_1)$ and $b_y = \sum_{i=n_1+1}^{n-n_2} y_i$. Fifth, set $g_i = \psi(u_i)$ and $h_i = \psi(v_i)$ with $\psi(u_i) = \max\left[-1, \min(1, u_i)\right]$ and $(v_i) = \max\left[-1, \min(1, v_i)\right]$ Then compute the percentage bend correlation through

$$r_{pbc} = \sum_{i=1}^{n} g_i h_i \Big/ \sqrt{\sum_{i=1}^{n} g_i^2 \sum_{i=1}^{n} h_i^2} \quad (13)$$

Bootstrap CIs

The bootstrap CIs I used in the simulation were constructed according to the following procedures (Efron & Tibshirani, 1993)[2]. First, I resampled each simulated dataset with replacement to form B number (e.g., 2,000) of bootstrap samples, each of them having the same sample size (n) as the original dataset. That is, Z^*(1st), Z^*(2nd),..., Z^*(Bth), where Z^* is a $n \times 2$ resample data matrix that consists of $n \times 1$ x_i and y_i, respectively.

Second, for each of the B bootstrap samples, I computed a bootstrap correlation r^*_{XY}, where r^*_{XY} can be 1 of the 12 correlation estimates evaluated in this study. Hence, B bootstrap correlations were computed for each of the 12 different estimators.

Third, given the B bootstrap r^*_{XY}s, I constructed the first type of bootstrap CI called the bootstrap standard interval (BSI), i.e.,

$$\text{BSI} = r_{XY} \pm \Phi^{-1}(1 - \alpha/2)\cdot \text{SD}(r^*_{XY}) \quad (14)$$

where r_{XY} is the observed correlation estimate in the original dataset (for each one of the 12 correlation estimators), $\Phi^{-1}(.)$ is the normal inverse cumulative distribution function, α is the Type I error (e.g., 5%), and $SD(r^*_{XY})$ is the SD of the B bootstrap correlations[3].

2) The n by 2 resample data matrix is obtained by resampling intact rows from the original data (bivariate bootstrapping) rather than resampling $X(i)$ and $Y(i)$ values independently (univariate bootstrapping). The former is appropriate for CI construction, whereas the latter might work better for testing a null hypothesis (Beasley et al., 2007). Given that the APA (2010, p. 34) stated that: "[b]ecause confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy", bivariate bootstrapping should have broader application than univariate bootstrapping that focuses on null-hypothesis testing.

However, BSI was found to be non-robust to asymmetric distributions because of the equal widths above and below the point estimate as a result of the $\pm\Phi(1-\alpha/2)\cdot SD(r^*_{XY})$ in (14). Thus, I also evaluated a second type of bootstrap CI called the bootstrap percentile interval (BPI), computed as

$$\text{BPI} = \left[r^*_{XY}(l), r^*_{XY}(u)\right] \tag{15}$$

where $l = \alpha/2$ and $u = (1-\alpha/2)$ with α as the Type I error (e.g., 5%), and hence, $(1-\alpha)$% is the CI constructed for a correlation estimate. The lower and upper limits of BPI depend on the ranks of the bootstrap correlations, which allow unequal widths above and below the point estimate. Thus, the BPI is found to be more robust to asymmetric (non-normal) distributions.

A third type of bootstrap CI is the bootstrap bias-corrected-and-accelerated CI (BCaI). It has been shown to improve the accuracy of the lower and upper limits in BPI especially when the distribution of bootstrap r^*_{XY}s is highly non-normal, e.g., highly skewed (Chan & Chan, 2004). Two correction factors, *i*, and *j*, are required to correct for the bias of BPI. The first factor *i*, is used to correct for the overall bias of the bootstrap r^*_{XY}s that deviate from the original correlation estimate (r_{XY}). That is,

$$i = \Phi^{-1}\left\{\#\left[r^*_{XY}(b) < r_{XY}\right]/B\right\} \tag{16}$$

where Φ^{-1} is the normal inverse cumulative function distribution, and $\#\left[r^*_{XY}(b) < r_{XY}\right]$ is the count function that counts the number of the bootstrap r^*_{XY}s below r_{XY} in the original dataset. The second correction factor (*j*) adjusts for the rate of change of the error of r_{XY} with respect to its true parameter value, which can be expressed as

$$j = \sum_{k=1}^{K}\left[r_{XY}(.) - r_{XY}(k)\right]^3 / 6\left\{\sum_{k=1}^{K}\left[r_{XY}(.) - r_{XY}(k)\right]^2\right\}^{3/2} \tag{17}$$

where $r_{XY}(k)$ is the jackknife value of the correlation r_{XY} obtained by removing the *k*th row of the original dataset $\boldsymbol{Z}$, and $r_{XY}(.)$ is the mean of the *n* jackknife estimates. Consequently, the BCaI can be estimated by

$$\text{BCaI} = \left[r^*_{XY}(B\cdot\alpha_1), r^*_{XY}(B\cdot\alpha_2)\right] \tag{18}$$

where $\alpha_1 = \Phi\left\{i + \frac{i + z_{1-(\alpha/2)}}{1 - j\left[i + z_{1-(\alpha/2)}\right]}\right\}$, and $\alpha_2 = \Phi\left\{i + \frac{i - z_{1-(\alpha/2)}}{1 - j\left[i - z_{1-(\alpha/2)}\right]}\right\}$.

3) A BSI could extend beyond the range of possible correlations (i.e., below -1 or above +1).

Monte Carlo Study

Design

Factor 1: Population Bivariate Linearity (θ)

Four levels— .10, .30, .50, and .80—were evaluated. The first three levels (.10, .30, .50) generally refer to a small, medium, and large effect size, respectively, which are commonly found in behavioral and social science research (Cohen, 1988). In addition, an extremely large value (.80) was also included to examine the impact of a very large effect on the correlation estimates.

Factor 2: Sample Size (n)

Three levels—50, 100, and 200—were examined, which correspond to relatively small, moderate, and large sample sizes frequently used in simulation studies (Li et al., 2011).

Factor 3: Proportion and SD (ζ, ϱ)

A total of six conditions that correspond to 3 different proportions of outliers and/or leverage points with 2 different *SD* units were examined. An ideal condition (i.e., no outliers/leverage points) was also examined, i.e., x and y are normally distributed, and are linearly related with a level of θ. In equation,

$$y = \theta x + e \tag{19}$$

where $x \sim N(0, 1)$, and e is the error term such that $e \sim N(0, \sigma_e)$, where $\sigma_e = \sqrt{1 - \theta^2}$.

The ideal condition in Equation (19) can be violated in behavioral and social sciences. Often, some of the x and/or y observations deviate from the general pattern of the remaining observations. According to previous simulation studies, (e.g., Algina et al., 2005), it is common that 10% of the x and/or y scores are these deviant O-LO with an extreme distance (i.e., 10 times of the *SD* in original scores) relative to the remaining scores. In this study, the 10% and 10 *SD*-unit is denoted as (10%, 10-*SD*). To provide a more complete picture about how these two factors influence the accuracy of r and the 11 robust correlation estimates, this study also includes (5%, 5-*SD*), (5%, 10-*SD*), (10%, 5-*SD*), (20%, 5-*SD*), and (20%, 10-*SD*), for a total of six different metric-based conditions.

The method used in the current simulation is based on the fact that a 5%, 10% or 20% random sample of the original x and/or y scores are multiplied by 5 or 10. Given that there is an equal chance that a portion (5%, 10%, or 20%) of both positive and negative x and/or y scores can be manipulated as outliers/leverage points, and x and y follow symmetric distributions, a good correlation estimate should be robust to these outliers and/or leverage points in a sample, and provide an accurate estimate of the true linear association between X and Y (θ). Given these metric-based conditions, there are four different cases of O-LO discussed below.

Factor 4: Outliers and Leverage Observations (Distribution)

Four cases are evaluated. For Case A (uniform y-outliers), the y-coordinate is such that the point is an outlier, and the point does not fit to the linear trend of most other points. For each level of Factor 3, ζ% of the data points were randomly selected and their y scores were multiple by ϱ. For Case B where the point is a leverage observation, ζ% of the data points were randomly selected and both x and y in the same x-y pair are multiplied by ϱ. For Case C in which the x-coordinate of the point is an outlier and it does not fit to the linear trend of most other points, ζ% of the data points were randomly selected and their x scores were multiple by ϱ. For Case D where both the x-coordinate and y-coordinate of the point are outliers and the x and y scores of the outlier points are correlated in the opposite direction of θ, ζ% of the x-y points were generated from Equation (19) where θ was replaced with -θ. For each of these data points both the x and y in the same x-y pair were multiplied by ϱ to become O-LO points.

In sum, the four factors were combined to produce a design with $4 \times 3 \times 6 \times 4 = 288$ conditions that contain outliers and/or leverage points (O-LO). In addition, a total of $4 \times 3 = 12$ ideal conditions in which x and y follow a normal distribution without O-LO are also evaluated. For each of the 300 conditions, 1,000 datasets were generated by first generating X scores as a random sample from a normal distribution with mean 0 and $SD = 1$ and e scores as a random sample from a normal distribution with mean 0 and $SD = \sqrt{1 - \theta^2}$. Corresponding Y scores were generated through Equation (19) such that the expected linear relationship is θ (Chan & Chan, 2004). Each of these 1,000 datasets was resampled $B = 2{,}000$ times to construct the 95% BSI, BPI, and BCaI surrounding each of the 12 correlation estimates. The code was written and executed in the R Project programming environment (R Core Team, 2017), and is available in the Supplementary Materials (Part A).

Evaluation Criteria

Two criteria were used to evaluate the performance of r and the 11 robust correlations. First, percentage bias was used to evaluate the deviation between the observed correlation estimate and the population measure of association θ, i.e., bias $= \left[\left(\overline{\delta_{XY}} - \theta\right)/\theta\right]$, where $\overline{\delta_{XY}}$ is the mean of the 1,000 sample correlation estimates (expressed as 1 of the 12 estimates under study). A correlation estimate is considered reasonable if the bias is within $\pm.10$ ((Li et al., 2011). The bias enables evaluation of the performance of a correlation estimate in a single condition. Second, for summarizing overall performance across T number of conditions, the mean-absolute-percentage-bias (MAPE) was used, MAPE $= \left(\sum_{t=1}^{T} |\text{bias}(t)|\right)/T$. A MAPE smaller than .10 is considered reasonable (Li et al., 2011).

Regarding the performance of the BSI, BPI, and BCaI, given that 95% CIs were constructed, the coverage was expected to be 950 out of 1,000 replications [or coverage probability (CP) = .95]. However, it is impossible for one to obtain a perfect CP of .95 in

the presence of sampling error. Thus, an observed CP that falls within the range (.925, .975) is considered acceptable (Chan & Chan, 2004).

Results

Estimates

The results showed that the presence of O-LO was the most influential factor affecting the performance of *r* and the 11 robust correlations as well as the associated bootstrap CIs. Hence, the findings are presented according to the four types of O-LO plus the conditions without O-LO. For the overall patterns of correlation estimates and bootstrap CIs across 300 simulation conditions, please refer to Figure A1 and Figure A2 in the Appendix, respectively.

Normality

As predicted, *r* produced highly accurate results with normal data (see Table 1). The biases ranged from -.017 to .013 with a mean of .000 [range-of-biases = (-.017, .013); mean-of-biases = .000]. All of the 12 conditions without OL-O yielded a bias within ±.10. MAPE was also appropriate (.007). The remaining 11 robust correlation estimates were reasonable: 10 of them (r_{CME}, r_{MAD}, r_{MED}, r_{bm}, r_Q, r_τ, r_s, r_{wLS}, r_{TRIM}, and, r_{pbc}) produced a mean bias within ±.10 and a MAPE smaller than .10. The only exception is r_w with a mean bias of -.108 and a MAPE of .108. Hence, the majority of the robust correlations can reasonably estimate the level of linearity θ when the assumption of normality and no O-LO are present in the data.

Case A (Y-Outliers)

The performance of *r* substantially decreased when *y*-outliers were present in the data, with range-of-biases = (-.495, -.124) and mean-of-biases = -.305, showing a substantial downward bias. None of the 72 conditions produced an acceptable bias. Accordingly, MAPE was also less than optimal (.305). The 11 robust correlations noticeably outperformed *r*. Ten of them (r_{CME}, r_{MAD}, r_{MED}, r_{bm}, r_Q, r_τ, r_s, r_{wLS}, r_{TRIM}, and, r_{pbc}) produced a mean bias within ±.10 and a MAPE smaller than .10, whereas r_w resulted in a slightly larger mean bias of -.122 and a MAPE of .122. Comparatively, r_Q appears to be the most accurate with a mean bias of -.018 and a MAPE of .026 under Case A. However, r_Q did not perform well under Cases C and D. Rather, r_{MAD}, r_{MED}, and r_{TRIM} performed best in these cases, and hence, they are further discussed in the following sections.

Case B (LO)

As predicted, *r* is reasonably robust to LO. The biases ranged from -.159 to .033 with a mean of -.027. Of the 72 conditions, 69 (or 95.8%) produced a bias within ±.10, indicating

Table 1

Percentage Biases of Correlation Estimates Under Normality and Cases A to D

Distribution	Statistic	Correlation											
		r	r_{CME}	r_{MAD}	r_{MED}	r_{bm}	r_{Q}	r_{τ}	r_{s}	r_{wLS}	r_{TRIM}	r_{w}	r_{pbc}
Normality	Mean	**.000**	**-.040**	**-.042**	**-.049**	**-.004**	**-.012**	**-.001**	**-.040**	**.029**	**-.060**	-.108	**-.045**
	SD	.009	.013	.061	.058	.012	.021	.011	.014	.016	.053	.032	.015
	Min	-.017	-.065	-.224	-.210	-.030	-.055	-.022	-.065	.006	-.189	-.165	-.066
	Max	.013	-.015	.006	.018	.018	.023	.028	-.019	.052	.019	-.058	-.024
	% ±.10	1.000	1.000	.917	.917	1.000	1.000	1.000	1.000	1.000	.833	.333	1.000
	MAPE	**.007**	**.040**	**.043**	**.052**	**.010**	**.019**	**.007**	**.040**	**.029**	**.064**	.108	**.045**
Case A	Mean	-.305	**-.095**	**-.064**	**-.072**	**.044**	**-.018**	**-.045**	**-.083**	**-.009**	**-.086**	-.122	**-.084**
	SD	.097	.099	.039	.036	.056	.031	.025	.028	.058	.044	.033	.031
	Min	-.495	-.333	-.197	-.205	-.060	-.149	-.115	-.160	-.220	-.250	-.208	-.157
	Max	-.124	.000	-.003	-.015	.186	.102	.012	-.034	.079	-.012	-.063	-.033
	% ±.10	.000	.681	.861	.806	.847	.972	.986	.750	.944	.694	.250	.694
	MAPE	.305	**.095**	**.064**	**.072**	**.052**	**.026**	**.045**	**.083**	**.036**	**.086**	.122	**.084**
Case B	Mean	**-.027**	-.123	**-.027**	**-.040**	.170	**-.014**	**-.004**	**-.088**	.299	**-.049**	-.124	**-.086**
	SD	.033	.092	.035	.039	.135	.026	.015	.031	.217	.039	.031	.034
	Min	-.159	-.431	-.171	-.212	-.003	-.119	-.066	-.197	.045	-.202	-.231	-.197
	Max	.033	-.034	.104	.041	.539	.070	.056	-.034	1.093	.045	-.063	-.033
	% ±.10	.958	.556	.958	.958	.375	.986	1.000	.681	.125	.931	.222	.722
	MAPE	**.029**	.123	**.032**	**.042**	.170	**.021**	**.010**	**.088**	.299	**.051**	.124	**.086**
Case C	Mean	-.776	-.326	**-.047**	**-.063**	**.095**	-.120	-.199	-.294	-.331	**-.069**	-.258	-.286
	SD	.160	.224	.039	.042	.070	.070	.111	.130	.193	.043	.090	.145
	Min	-.987	-.798	-.187	-.213	-.054	-.354	-.416	-.519	-.673	-.176	-.430	-.543
	Max	-.389	-.104	.055	.036	.245	-.012	-.061	-.127	-.030	.025	-.123	-.110
	% ±.10	.000	.000	.917	.847	.569	.528	.292	.000	.167	.792	.000	.000
	MAPE	.776	.326	**.050**	**.064**	**.097**	.120	.199	.294	.331	**.070**	.258	.286
Case D	Mean	-1.536	-.533	**-.071**	**-.091**	**.032**	-.236	-.402	-.498	-1.170	**-.094**	-.390	-.482
	SD	.313	.382	.049	.053	.097	.130	.219	.233	.592	.048	.161	.259
	Min	-1.984	-1.433	-.210	-.225	-.269	-.553	-.751	-.844	-2.436	-.226	-.639	-.911
	Max	-.789	-.167	.024	.003	.289	-.003	-.121	-.206	-.199	-.014	-.174	-.178
	% ±.10	.000	.000	.764	.653	.694	.125	.000	.000	.000	.611	.000	.000
	MAPE	1.536	.533	**.072**	**.091**	**.076**	.236	.402	.498	1.170	**.094**	.390	.482

Note: r is Pearson's correlation, r_{CME} is median estimator correlation, r_{MAD} is median absolute deviation correlation, r_{MED} is median correlation coefficient, r_{bm} is biweight midcorrelation, r_{Q} is quadrant correlation coefficient, r_{τ} is Kendall's τ correlation, r_{s} is Spearman's correlation, r_{wLS} is weighted least-squares correlation, r_{TRIM} is trimmed correlation, r_{W} is Winsorized correlation, r_{pbc} is percentage bend correlation, "% ±.10" indicates the proportion of conditions with a percentage bias within the criterion of ±.10, and MAPE is defined as the mean absolute percentage error. The results are presented in bold when MAPE is less than .10, and when Mean of bias is less than .10.

a reasonable fit. MAPE was also acceptable (.029). Of the 11 robust correlations, only 7 (r_{MAD}, r_{MED}, r_{Q}, r_{τ}, r_{s}, r_{TRIM}, and, r_{pbc}) produced a bias within ±.10. As noted above, r_{MAD},

r_{MED}, and r_{TRIM} also behave appropriately under Cases C and D, and hence, they are further discussed. For r_{MAD}, the range-of-biases (-.171, .104) and the mean-of-biases was 027. Of the 72 conditions, 69 (or 95.8%) produced a bias within ±.10. MAPE was also good (.032). For r_{MED}, the range-of-biases was (-.212, .041) and mean-of-biases was -.040. Of the 72 conditions, 69 (or 95.8%) yielded a bias within ±.10. MAPE is also appropriate (.042). For r_{TRIM}, the mean-of-biases was (-.202, .045) and mean-of-biases was -.049. Of the 72 conditions, 67 (93.1%) produced a bias within ±.10. MAPE is also acceptable (.051).

Case C (X-Outliers)

As in Case A, *r* proved to be a poorly performing estimate of the bivariate linear relationships under Case C conditions. The range-of-biases was (-.987, -.389) and mean-of-biases was -.776, showing a severe-downward bias. Of the 72 conditions, 0 (or 0%) produced an acceptable bias. MAPE was also very poor (.776). Regarding the 11 robust correlations, only 3 (r_{MAD}, r_{MED}, and r_{TRIM}) yielded an acceptable bias within ±.10. For r_{MAD}, the range-of-biases was (-.187, .055) and mean-of-biases was -.047. Of the 72 conditions, 66 (or 91.7%) produced a bias within ±.10. MAPE was also appropriate (.050). For r_{MED}, the range-of-biases was (-.213, .036) and mean-of-biases was -.063. Of the 72 conditions, 61 (or 84.7%) yielded a bias within ±.10. MAPE was also reasonable (.064). For r_{TRIM}, the range-of-biases was (-.176, .025) and mean-of-biases was -.069. Of the 72 conditions, 57 (79.2%) produced a bias within ±.10. MAPE was also reasonable (.070).

Case D (LO with Opposite Correlation)

Unsurprisingly, Case D was found to be the data condition most detrimental to the accuracy of correlation estimates. For *r*, the range-of-biases was (-1.984, -.789) and mean-of-biases was -1.536. None of the 72 conditions produced an acceptable bias. MAPE was also high (1.536) indicating a poor fit. Comparing the 11 robust correlations, r_{MAD}, r_{MED}, and r_{TRIM} produced reasonable results. For r_{MAD}, the range-of-biases was (-.210, -.024) and mean-of-biases was -.071. Of the 72 conditions, 55 (or 76.4%) produced a bias within ±.10. MAPE was also reasonable (.072). For r_{MED}, the range-of-biases was (-.225, .003) and mean-of-biases was -.091. Of the 72 conditions, 47 (or 65.3%) yielded a bias within ±.10. MAPE was also reasonable (.091). For r_{TRIM}, the range-of-biases was (-.226, .014) and mean-of-biases was -.094. Of the 72 conditions, 44 (61.1%) produced a bias within ±.10. MAPE was also reasonable (.094).

Bootstrap CIs

Normality

The 3 bootstrap CIs constructed for *r* performed acceptably, with a mean CP of .937 for BSI, .942 for BPI, and .944 for BCaI (Table 2). Of the 12 conditions, 9 (or 75%) yielded an acceptable CP for BSI, and 11 (or 91.7%) resulted in an acceptable CP for both BPI

and BCaI. The remaining 11 robust correlations also produced good results, in general, with mean CPs ranging from .927 to .968. The bootstrap CIs constructed for the three key robust correlations, r_{MAD}, r_{MED}, and r_{TRIM}, were appropriate. For r_{MAD}, the BSI and BCaI were appropriate, with a mean CP of .968 and .943, respectively. For r_{MED}, the mean CPs were .965 and .932 for BSI and BCaI, respectively, which are appropriate. For r_{TRIM}, the mean coverage probabilities were .958, .967, and .939, respectively, for the BSI, BPI, and BCaI, which are also good.

Case A (Y-Outliers)

Given that r is not robust to Case A, the associated bootstrap CIs are also less than optimal. The mean CPs are only .583, .616, and .50 for BSI, BPI, and BCaI, respectively. Of the 72 conditions, only 1 (or 1.4%), 10 (or 13.9%), and 4 (or 5.6%) conditions based on BSI, BPI, and BCaI, respectively, yielded a CP within (.925, .975), indicating a poor fit. Comparatively, the 11 robust correlations yielded much more reasonable CPs, ranging from .886 to .967. Regarding the bootstrap CIs for r_{MAD}, one of the 3 robust correlation estimates as discussed above, the mean CPs were .967 and .931 for BSI, and BCaI respectively, showing acceptable fit. The BSI and BCaI for r_{MED}, was also good with mean CPs of .966 and .930, respectively. The CIs for the last of the acceptably performing robust correlations, r_{TRIM}, also performed well, with mean CPs of .960, .958, and .937 for BSI, BPI, and BCaI, respectively.

Case B (LO)

Although r was found to be reasonably robust to leverage observations, the associated BSI, BPI, and BCaI were less than optimal. The mean CPs were .856, .909, and .886 based on BSI, BPI, and BCaI, respectively. Of the 72 conditions, only 4.2% (BSI), 20.8% (BPI), and 19.4% (BCaI) of these conditions produced a CP within (.925, .975). Comparatively, the CPs yielded by BSI, BPI, and BCaI surrounding the 11 robust correlations (except r_{wLS}) were more reasonable. Specifically, the BSI, and BCaI constructed for r_{MAD} resulted in good mean CPs of .966 and .941, respectively. Of the 72 conditions, 81.9% (BSI) and 95.8% (BCaI) yielded a CPs within (.925, .975), respectively. Regarding r_{MAD}, BSI and BCaI resulted in good mean CPs of .968 and .933, respectively. Of the 72 conditions, 76.4% and 95.5% of them, respectively, produced a CP within (.925, .975). For r_{TRIM}, the mean CPs are .964, .974, and .945 for BSI, BPI, and BCaI, respectively, indicating an excellent fit. Of the 72 conditions, 75%, 55.6%, and 98.6% of conditions yielded a CP within [.925, .975].

Case C (X-Outliers)

Similar to Case A, the bootstrap CIs surrounding r were not robust to x-outliers. The mean CPs were .588 for BSI, .641 for BPI, and .589 for BCaI, which are undesirable. Of the 72 conditions, 0%, 1.4%, and 1.4% resulted in a CP within (.925, .975) based on BSI, BPI, and BCaI, respectively. Comparatively, the bootstrap CIs constructed for the robust

Table 2

Coverage Probabilities of the Bootstrap CIs (i.e., BSI, BPI, and BCaI) under Normality and Cases A to D

			Correlation											
Distribution	**CI**	**Statistic**	***r***	r_{CME}	r_{MAD}	r_{MED}	r_{bm}	r_{Q}	r_{τ}	r_{s}	r_{wLS}	r_{TRIM}	r_{w}	r_{pbc}
Normality	BSI	Mean	**.936**	**.948**	**.968**	**.965**	**.940**	**.962**	**.944**	**.948**	**.927**	**.958**	**.928**	**.944**
		SD	.013	.013	.012	.013	.010	.010	.008	.011	.010	.019	.038	.008
		Min	.914	.935	.954	.950	.926	.945	.930	.930	.904	.935	.834	.931
		Max	.951	.977	.995	.994	.958	.981	.954	.969	.941	.993	.962	.954
		%	.833	.917	.833	.833	1.000	.917	1.000	1.000	.583	.750	.750	1.000
	BPI	Mean	**.942**	**.948**	.987	.985	**.941**	.983	**.952**	**.945**	**.943**	**.967**	.915	**.943**
		SD	.009	.013	.009	.007	.010	.010	.007	.008	.008	.008	.059	.011
		Min	.924	.917	.973	.972	.925	.964	.935	.932	.932	.956	.764	.921
		Max	.951	.964	1.000	.995	.961	.999	.960	.957	.957	.984	.966	.958
		%	.917	.917	.167	.083	1.000	.250	1.000	1.000	1.000	.833	.583	.917
	BCaI	Mean	**.943**	**.940**	**.943**	**.932**	**.939**	**.925**	**.956**	**.945**	**.940**	**.939**	.904	**.941**
		SD	.009	.013	.006	.011	.016	.011	.008	.009	.007	.007	.070	.017
		Min	.929	.909	.935	.913	.900	.903	.937	.928	.931	.927	.730	.906
		Max	.954	.957	.951	.949	.962	.938	.965	.959	.951	.953	.963	.961
		%	1.000	.917	1.000	.833	.833	.750	1.000	1.000	1.000	1.000	.583	.833
Case A	BSI	Mean	.583	.907	**.967**	**.966**	**.947**	**.961**	**.943**	**.926**	.923	**.960**	.917	**.926**
		SD	.350	.135	.011	.013	.013	.009	.036	.063	.099	.016	.064	.068
		Min	.000	.227	.945	.946	.884	.945	.686	.530	.315	.934	.617	.496
		Max	.936	.989	.991	.994	.967	.984	.976	.966	.974	.996	.970	.969
		%	.014	.736	.764	.806	.972	.917	.903	.833	.903	.819	.611	.847
	BPI	Mean	.616	.922	.981	.977	**.945**	.985	**.938**	.907	**.939**	**.958**	.898	.906
		SD	.352	.108	.015	.018	.014	.008	.051	.091	.096	.033	.092	.110
		Min	.000	.292	.912	.890	.854	.966	.604	.417	.259	.805	.498	.253
		Max	.941	.980	.997	.995	.962	.999	.969	.965	.981	.988	.974	.966
		%	.139	.750	.222	.306	.972	.111	.889	.694	.889	.681	.556	.722
	BCaI	Mean	.560	.900	**.936**	**.930**	**.930**	.923	**.935**	.905	.914	**.937**	.886	.900
		SD	.354	.123	.016	.015	.021	.012	.061	.095	.117	.018	.102	.113
		Min	.000	.274	.864	.857	.797	.897	.557	.408	.226	.855	.481	.267
		Max	.938	.967	.954	.957	.957	.947	.972	.967	.962	.963	.978	.966
		%	.056	.792	.861	.750	.778	.444	.861	.667	.847	.875	.486	.667
Case B	BSI	Mean	.856	**.931**	**.966**	**.968**	**.954**	**.960**	**.937**	**.932**	.745	**.964**	.919	**.939**
		SD	.051	.069	.010	.011	.039	.008	.011	.023	.193	.016	.056	.021
		Min	.693	.571	.947	.947	.687	.942	.911	.800	.038	.938	.698	.819
		Max	.932	.993	.989	.991	.994	.982	.958	.966	.940	.996	.971	.971
		%	.042	.708	.819	.764	.750	.958	.847	.819	.028	.750	.681	.875
	BPI	Mean	.909	**.934**	.986	.986	**.927**	.983	**.947**	**.926**	.847	**.974**	.904	**.932**
		SD	.016	.043	.007	.006	.071	.007	.008	.036	.167	.009	.077	.034
		Min	.870	.730	.966	.969	.460	.969	.929	.742	.099	.953	.613	.746
		Max	.943	.980	.998	.995	.970	.996	.966	.962	.954	.991	.969	.964
		%	.208	.750	.069	.056	.806	.181	1.000	.653	.417	.556	.583	.778
	BCaI	Mean	.886	.899	**.941**	**.933**	.889	.923	**.952**	.917	.773	**.945**	.891	.920
		SD	.039	.084	.008	.010	.087	.013	.008	.049	.167	.009	.093	.052
		Min	.790	.563	.921	.902	.371	.893	.935	.691	.130	.924	.557	.680
		Max	.943	.963	.958	.955	.949	.950	.969	.964	.917	.962	.972	.970
		%	.194	.597	.958	.847	.361	.472	1.000	.611	.000	.986	.514	.639

Table 2 (continued)

Distribution	CI	Statistic	Correlation											
			r	r_{CME}	r_{MAD}	r_{MED}	r_{bm}	r_{Q}	r_{τ}	r_{s}	r_{wLS}	r_{TRIM}	r_{w}	r_{pbc}
Case C	BSI	Mean	.588	.782	**.967**	**.967**	**.958**	**.944**	.844	.731	.879	**.964**	.756	.773
		SD	.270	.285	.011	.013	.020	.052	.195	.282	.058	.015	.270	.264
		Min	.000	.000	.946	.943	.869	.659	.035	.000	.684	.933	.001	.001
		Max	.881	.993	.992	.994	.995	.983	.968	.949	.963	.997	.964	.958
		%	.000	.417	.778	.806	.750	.792	.542	.319	.208	.764	.361	.375
	BPI	Mean	.644	.732	.985	.982	**.945**	**.955**	.818	.708	.915	**.968**	.719	.726
		SD	.310	.291	.008	.010	.022	.077	.223	.298	.088	.016	.295	.294
		Min	.000	.000	.954	.931	.804	.539	.017	.000	.435	.893	.000	.000
		Max	.925	.968	.996	.995	.973	.995	.958	.956	.970	.995	.971	.965
		%	.014	.333	.139	.153	.917	.333	.472	.278	.736	.694	.319	.319
	BCaI	Mean	.589	.707	**.938**	**.930**	.916	.870	.802	.683	.841	**.943**	.705	.700
		SD	.305	.308	.009	.008	.034	.090	.238	.314	.100	.010	.305	.312
		Min	.000	.000	.911	.908	.707	.451	.010	.000	.368	.914	.000	.000
		Max	.925	.962	.958	.949	.947	.949	.964	.963	.929	.967	.969	.973
		%	.014	.333	.972	.764	.444	.139	.458	.278	.028	.917	.306	.333
Case D	BSI	Mean	.359	.690	**.966**	**.965**	**.957**	.881	.676	.564	.628	**.959**	.625	.622
		SD	.323	.346	.015	.015	.020	.177	.333	.371	.313	.020	.351	.358
		Min	.000	.000	.884	.890	.879	.062	.000	.000	.000	.846	.000	.000
		Max	.822	.991	.993	.994	.993	.981	.979	.960	.973	.997	.962	.961
		%	.000	.361	.778	.764	.792	.653	.250	.167	.125	.778	.222	.264
	BPI	Mean	.444	.585	.977	**.972**	**.942**	.878	.648	.549	.701	**.952**	.590	.572
		SD	.368	.374	.028	.036	.029	.218	.348	.377	.327	.052	.368	.375
		Min	.000	.000	.778	.732	.777	.010	.000	.000	.000	.621	.000	.000
		Max	.910	.966	.996	.996	.979	.994	.962	.963	.970	.996	.971	.963
		%	.000	.236	.222	.319	.889	.319	.306	.194	.333	.625	.222	.208
	BCaI	Mean	.389	.572	**.932**	.923	.919	.784	.629	.529	.620	**.932**	.580	.553
		SD	.348	.377	.025	.022	.028	.204	.355	.382	.308	.029	.371	.382
		Min	.000	.000	.761	.780	.799	.028	.000	.000	.000	.743	.000	.000
		Max	.875	.963	.957	.946	.950	.929	.969	.965	.937	.956	.978	.970
		%	.000	.208	.847	.639	.556	.042	.278	.181	.028	.833	.222	.222

Note: r is Pearson's correlation, r_{CME} is median estimator correlation, r_{MAD} is median absolute deviation correlation, r_{MED} is median correlation coefficient, r_{bm} is biweight midcorrelation, r_{Q} is quadrant correlation coefficient, r_{τ} is Kendall's τ correlation, r_{s} is Spearman's correlation, r_{wLS} is weighted least-squares correlation, r_{TRIM} is trimmed correlation, r_{W} is Winsorized correlation, and r_{pbc} is percentage bend correlation, "%" indicates the proportion of conditions with a coverage probability within (.925, .975). The results are presented in bold when mean of coverage probabilities is within (.925, .975).

correlations (except r_{wLS}) were reasonable. Specifically, the BSI and BCaI constructed for r_{MAD} (or r_{MED}) are good, with mean CPs of .967 and .938 (or .967 and .930) respectively. Of the 72 conditions, 77.8% and 97.2% (or 80.6% and 76.4%) of the BSIs and BCaIs, respectively, produced a CP within (.925, .975). Regarding r_{TRIM}, the mean CPs were .964,

.968, and .943 for BSI, BPI, and BCaI, respectively. Of the 72 conditions, 76.4% (BSI), 69.4% (BPI), and 91.7% (BCaI) produced a CP within (.925, .975).

Case D (LO With Opposite Correlation)

r is not robust to Case D. The mean CPs were .359, .444, and .389 for BSI, BPI, and BCaI, respectively. Of the 72 conditions, 0% from BSI, BPI, or BCaI fell within the range of (.925, .975). Comparatively, the bootstrap CIs for r_{MAD}, r_{MED}, and r_{TRIM} were appropriate. For r_{MAD}, the mean CPs were .966 and.932 for BSI and BCaI, respectively. Of the 72 conditions, 77.8% and 84.7% yielded a CP within (.925, .975). For r_{MED}, the mean CPs were .965 and .923 for BSI and BCaI, respectively. Of the 72 conditions, 76.4% and 63.9% yielded a CP within (.925, .975). The bootstrap CIs for r_{TRIM} appear to have performed the best. The mean CPs were .959, .952, and .932 for BSI, BPI, and BCaI, respectively. Of the 72 conditions, 77.8%, 62.5%, and 83.3% yielded a CP within (.925, .975).

Conclusion and Discussion

In behavioral and social science research, data may contain O-LO in practice. The results of my simulation study suggest that r_{MAD} and r_{MED} (with BSI and BCaI) and r_{TRIM} (with BSI, BPI, and BCaI) are good alternatives to consider. I recommend that researchers compute and report these robust correlation estimates in addition to the traditional r when their data contain O-LO. Under these conditions, it would be reasonable to draw inferences based on the recommended robust CIs and include r only as a supplementary analysis. The R function (robr) provided in the Supplementary Materials (Part B) enables researchers to easily implement these recommendations. In addition, the Supplementary Materials (Part B) provide a real-world example to illustrate the differences between Pearson's and robust correlations and the implications of using robust estimates of bivariate linearity.

Given the more desirable performance of r_{MAD}, r_{MED}, and r_{TRIM} over r when data contain O-LO, one may ask whether or not we can use these robust correlations routinely. When O-LO exist, the answer provided by the results of this study is crystal clear. Yes, because the robust alternatives are more accurate than r under these conditions. When data do not contain O-LO, r is often regarded as more efficient (e.g., more powerful, narrower CI) than the robust correlations. In the past, this claim may have contributed to resistance to the broader use of robust statistics. Pernet et al. (2013) have compared the power rates of r and its robust alternatives, and concluded that these alternatives "provide accurate estimates of the true correlations for normal and contaminated data with no or minimal loss of power and adequate control over the false positive rate" (p. 11), especially with small correlation values and small sample sizes ($\rho < .2$, $n < 150$), and large correlation values and large sample sizes ($\rho > .3$, $n > 250$). For other combinations

of correlation values and sample sizes, r is slightly more powerful than r_s (maximum 10%).

The present simulation also found similar patterns of results, but the benefits of r are not obvious, given that when data do not include O-LO, the mean biases, the number of conditions that produced a bias within ±.10, the mean coverage probabilities, and the number of conditions that produced a CP within (.925, .975) are highly comparable between r and the 11 robust correlations. Given that the call for new statistical practices that focus on effect size and CI instead of p value in psychological research (e.g., Cumming, 2014), the only benefit (i.e., a possibly slightly more powerful test) from r may diminish in importance. This is especially probable considering that a slight reduction in power can be avoided by a slight increase in sample size, but collecting O-LO data within a sample cannot be so easily avoided.

Combining Pernet et al.'s (2013) results with the findings in this study, the benefits of these robust alternatives to r are apparent: They are more robust than r with O-LO that are commonly found in behavioral and social data. There seems to be no justification for continuing to rely almost exclusively on an estimator that is likely to be biased under many common conditions when there are viable robust alternatives. As behavioral and social science researchers we have everything to gain and little to lose by changing our practices to increase the accuracy with which we estimate the magnitude of bivariate linear relationships. The robust correlation estimates identified in this study, r_{MAD}, r_{MED}, and r_{TRIM}, should be considered and even routinely used by researchers in the future.

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.

Supplementary Materials

The supplementary materials provide in Part A the simulation code used in the present study. Part B provide a real-world example based on Mychasiuk's (2017) behavioral data of 74 adolescent rats to illustrate the differences between Pearson's and robust correlations and the implications of using robust estimates of bivariate linearity. Using the R Project statistical software (R Core Team, 2017) researchers can use the "robr" function to compute r and each of the robust correlation estimates, together with their bootstrap CIs (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Li, J. C. (2022). *Supplementary materials to "Bootstrap confidence intervals for 11 robust correlations in the presence of outliers and leverage observations"* [Simulation code, application example]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.7051

References

Albers, M. J. (2017). *Introduction to quantitative data analysis in the behavioral and social sciences.* John Wiley & Sons.

Algina, J., Keselman, H. J., & Penfield, R. D. P. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods, 10*(3), 317–328. https://doi.org/10.1037/1082-989X.10.3.317

APA. (2010). *Publication manual of the American Psychological Association* (6th ed.). American Psychological Association.

Beasley, W. H., DeShea, L., Toothajer, L., Mendoza, J., Bard, D., & Rodgers, J. (2007). Boostrapping to test for nonzero population correlation coefficients using univariate sampling. *Psychological Methods, 12*(4), 414–433. https://doi.org/10.1037/1082-989X.12.4.414

Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics, 21*(4), 593–600. https://doi.org/10.1214/aoms/1177729754

Chan, W., & Chan, D. W.-L. (2004). Bootstrap standard error and confidence intervals for the correlation corrected for range restriction: A simulation study. *Psychological Methods, 9*(3), 369–385. https://doi.org/10.1037/1082-989X.9.3.369

Chen, Y.-J. (2006). *Robust properties of generalized correlation coefficients, with applications to cross-over designs* [Unpublished doctoral dissertation]. Pennsylvania State University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Taylor & Francis Group.

Copas, J. (1988). Binary regression models for contaminated data. *Journal of the Royal Statistical Society: Series B (Methodological), 50*(2), 225–253. https://doi.org/10.1111/j.2517-6161.1988.tb01723.x

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science, 25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Devlin, S. J., Gnanadesikan, R., & Kettenring, J. R. (1975). Robust estimation and outlier detection with correlation coefficient. *Biometrika, 62*(3), 531–545. https://doi.org/10.1093/biomet/62.3.531

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap.* Chapman & Hall/CRC.

Falk, M. (1998). A note on the correlation median for elliptical distributions. *Journal of Multivariate Analysis, 67*(2), 306–317. https://doi.org/10.1006/jmva.1998.1775

Gnanadesikan, R., & Kettenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics, 28*(1), 81–124. https://doi.org/10.2307/2528963

Huber, P. J. (1981). *Robust statistics.* John Wiley & Sons.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika, 30*(1-2), 81–93. https://doi.org/10.1093/biomet/30.1-2.81

Kendall, M., & Stuart, A. (1977). *The advanced theory of statistics* (4th ed.). Macmillan.

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

Li, J. C.-H., Chan, W., & Cui, Y. (2011). Bootstrap standard error and confidence intervals for the correlations corrected for indirect range restriction. *British Journal of Mathematical & Statistical Psychology, 64*(3), 367–387. https://doi.org/10.1348/2044-8317.002007

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Mychasiuk, R. (2017). *Behavioral and pathophysiological outcomes associated with caffeine consumption and repetitive mild traumatic brain injury (RmTBI) in adolescent rats (Version 1.0)* [Data set]. Scholars Portal Dataverse. https://doi.org/10.5683/SP/8RODEV

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

Pasman, V. R., & Shevlyakov, G. L. (1987). Robust methods of estimation of correlation coefficients. *Automation and Remote Control, 48*, 332–340. https://doi.org/10.1134/S0005117906120071

Pernet, C. R., Wilcox, R., & Rousselet, G. (2013). Robust correlation analyses: False positive and power validation using a new open source Matlab toolbox. *Frontiers in Psychology, 3*, Article e606. Advance online publication. https://doi.org/10.3389/fpsyg.2012.00606

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org/.

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

Ruscio, J. (2008). Constructing confidence intervals for Spearman's rank order correlation with ordinal data. *Journal of Modern Applied Statistical Methods, 7*(2), 416–434. https://doi.org/10.22237/jmasm/1225512360

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47*(2), 201–223. https://doi.org/10.1080/00273171.2012.658329

Shevlyakov, G., & Smirnov, P. (2011). Robust estimation of the correlation coefficient: An attempt of survey. *Austrian Journal of Statistics, 40*(1–2), 147–156.

Shevlyakov, G. L., & Vilchevsky, N. O. (2002). Minimax variance estimation of a correlation coefficient for epsilon-contaminated bivariate normal distributions. *Statistics & Probability Letters, 57*(1), 91–100. https://doi.org/10.1016/S0167-7152(02)00058-5

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology, 15*(1), 72–101. https://doi.org/10.2307/1412159

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics* (pp. 448–485). Stanford University Press.

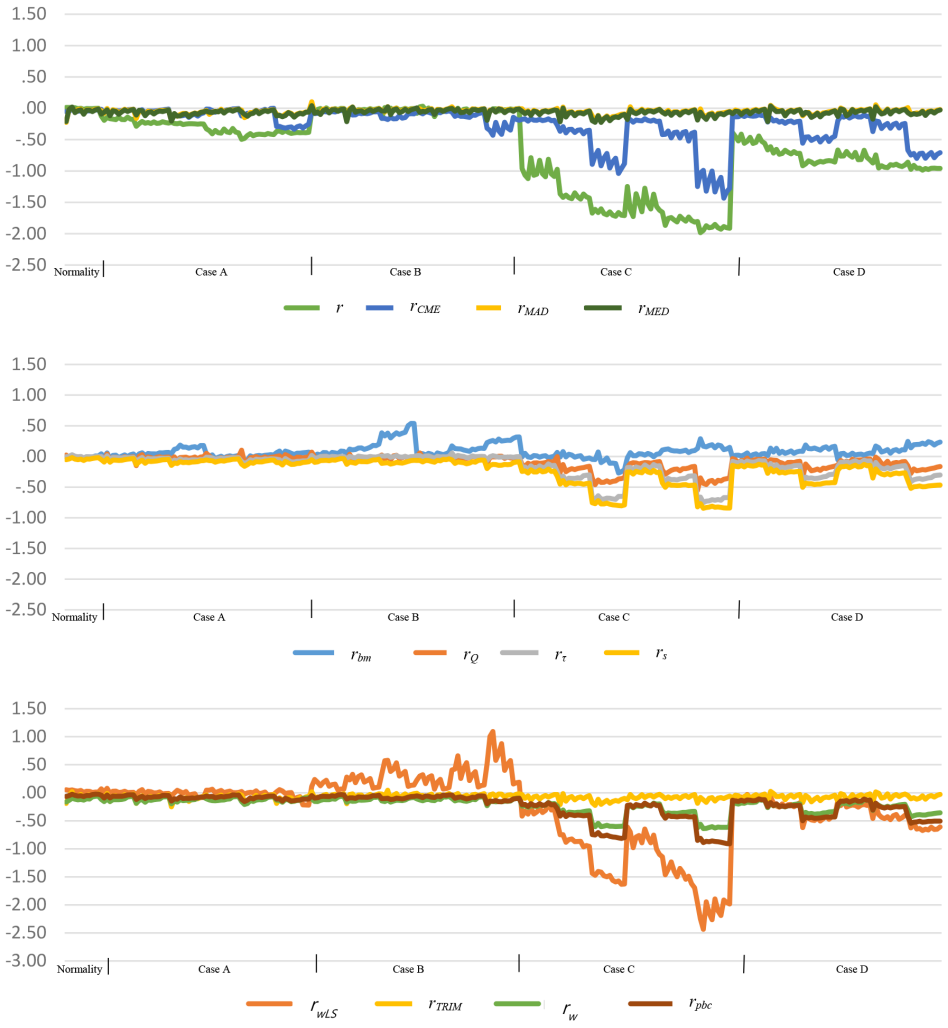
Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Elsevier.

Yuan, K.-H., & Zhong, X. (2013). Robustness of fit indices to outliers and leverage observations in structural equation modeling. *Psychological Methods, 18*(2), 121–136. https://doi.org/10.1037/a0031604

Appendix

Figure A1

Percentage Biases of Correlation Estimates Across 300 Conditions



Note. r is Pearson's correlation, r_{CME} is median estimator correlation, r_{MAD} is median absolute deviation correlation, r_{MED} is median correlation coefficient, r_{bm} is biweight midcorrelation, r_Q is quadrant correlation coefficient, r_τ is Kendall's τ correlation, r_s is Spearman's correlation, r_{wLS} is weighted least-squares correlation, r_{TRIM} is trimmed correlation, r_W is Winsorized correlation, and r_{pbc} is percentage bend correlation.

Figure A2

Coverage Probabilities of BSI, BPI, and BCaI for r and 11 Robust Correlations Across 300 Conditions

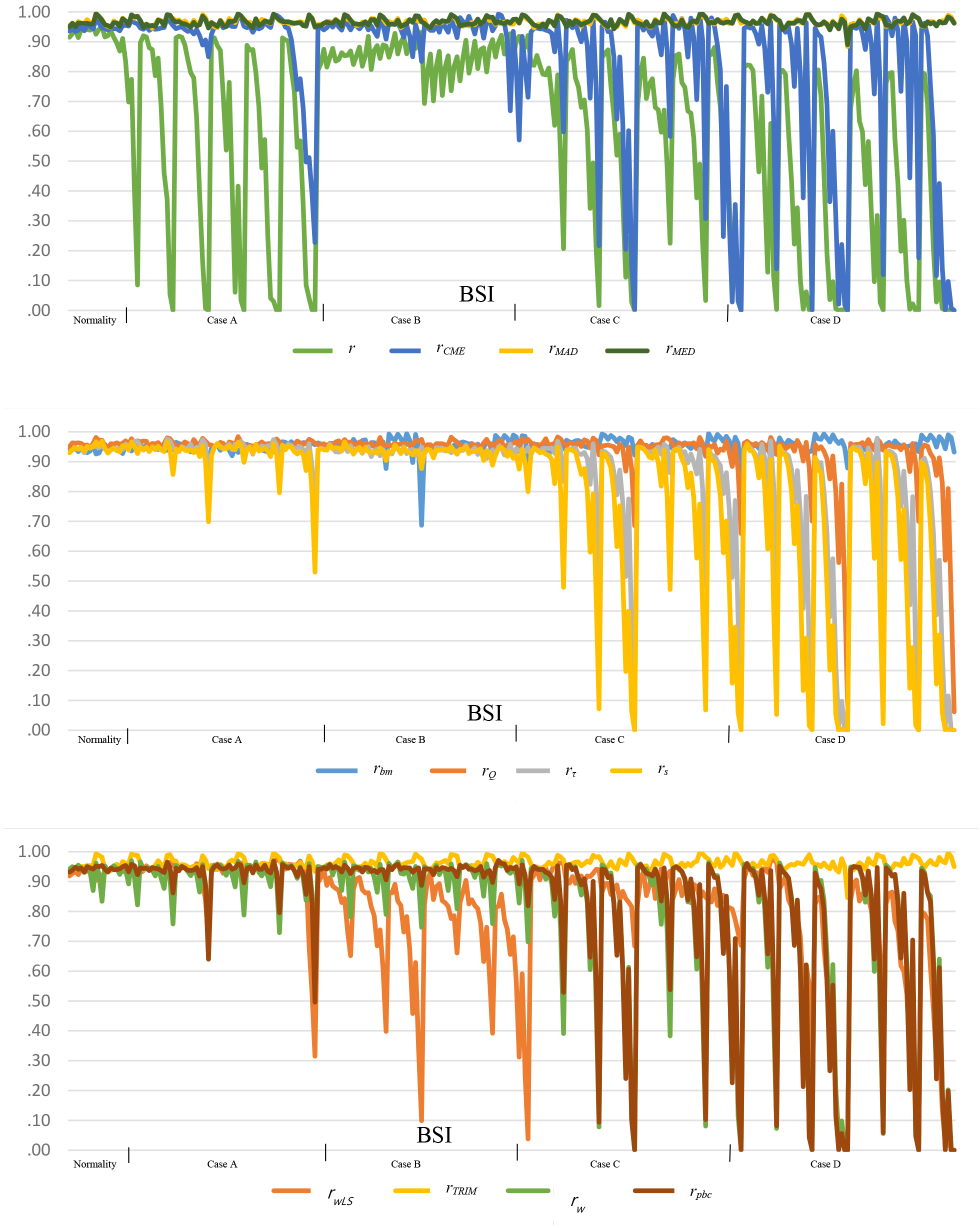


Figure A2 (continued)

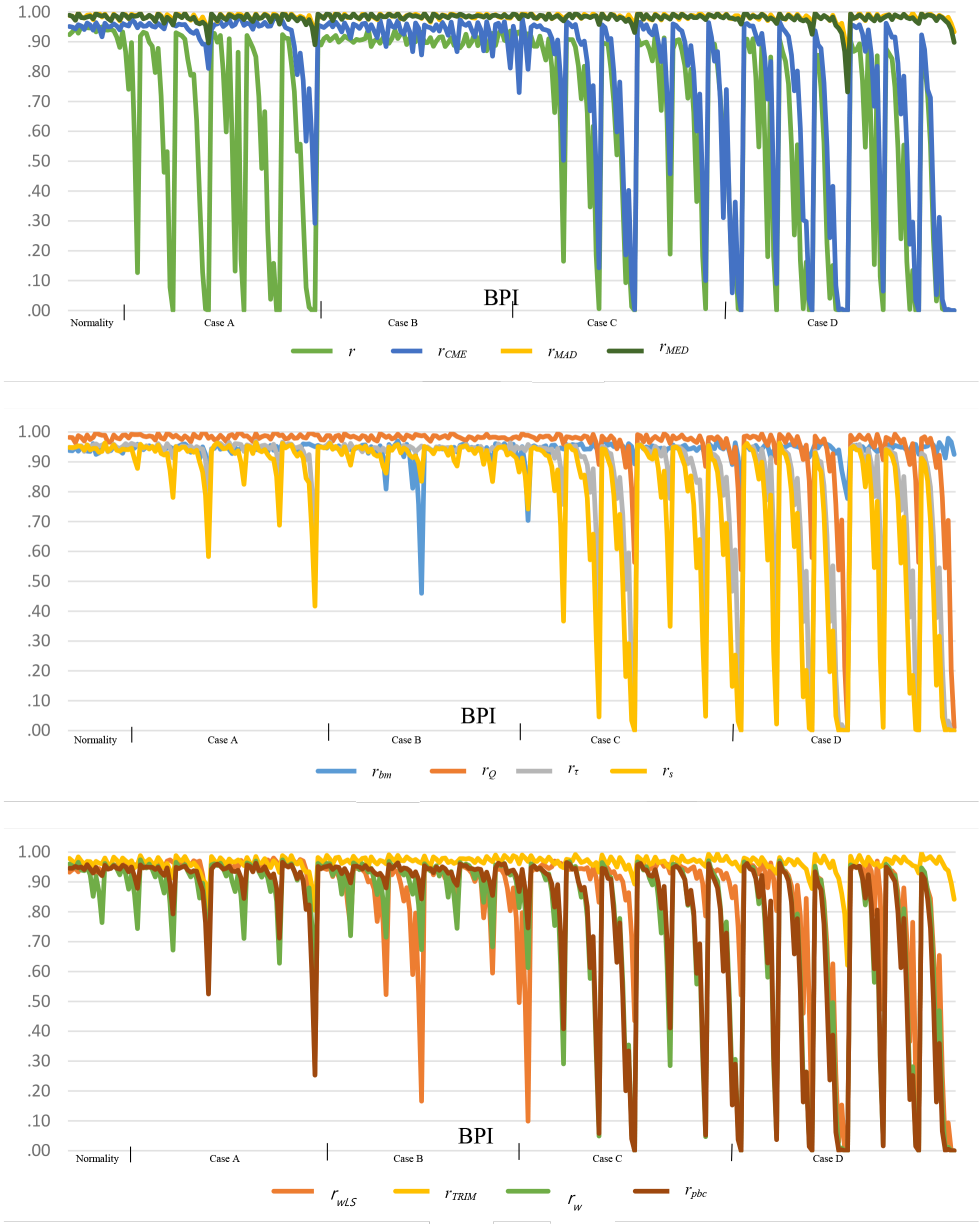
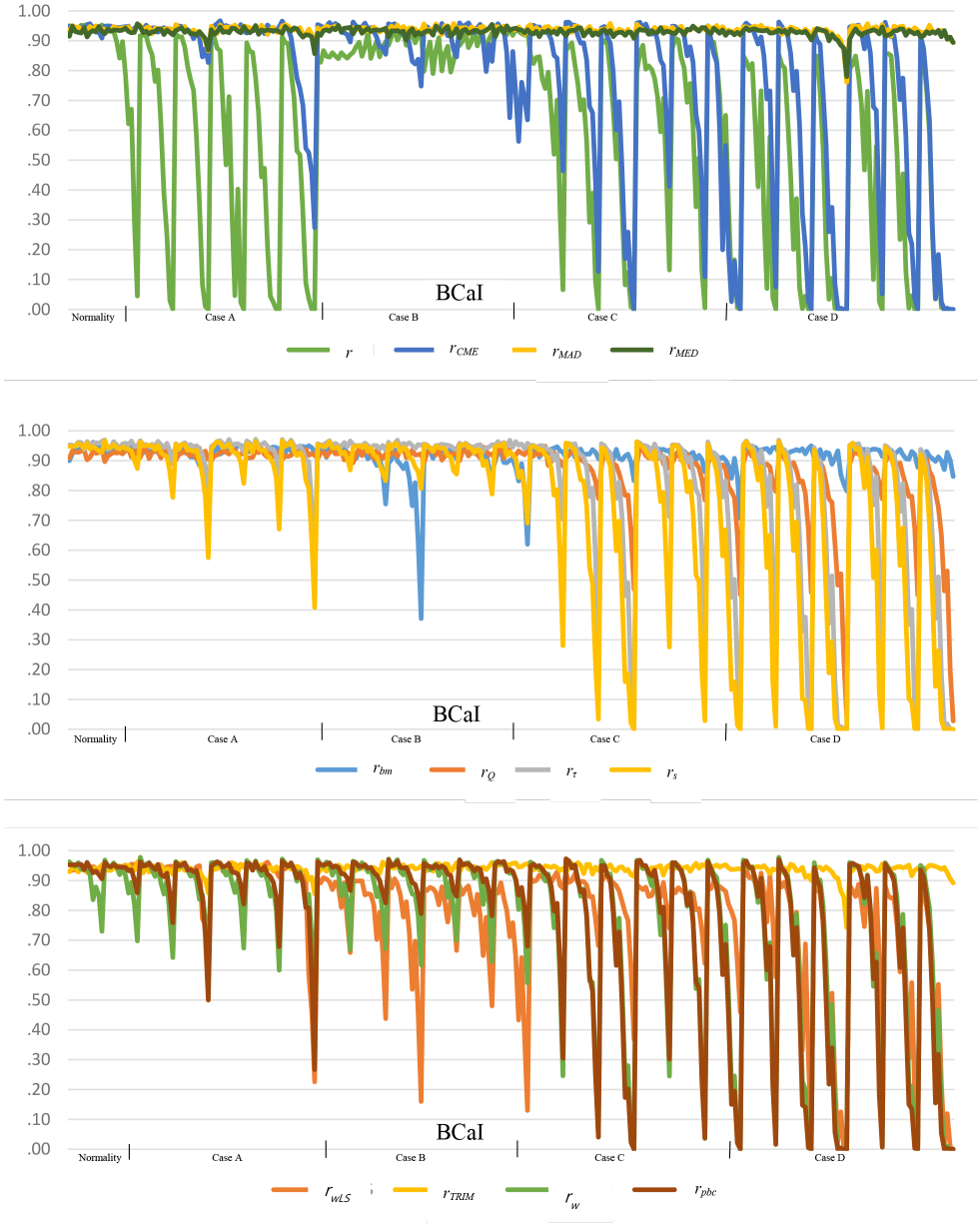


Figure A2 (continued)



Note. r is Pearson's correlation, r_{CME} is median estimator correlation, r_{MAD} is median absolute deviation correlation, r_{MED} is median correlation coefficient, r_{bm} is biweight midcorrelation, r_Q is quadrant correlation coefficient, r_τ is Kendall's τ correlation, r_s is Spearman's correlation, r_{wLS} is weighted least-squares correlation, r_{TRIM} is trimmed correlation, r_W is Winsorized correlation, and r_{pbc} is percentage bend correlation.

Methodology
2022, Vol. 18(2), 99–125
https://doi.org/10.5964/meth.8467