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

PsychOpen GOLD

Original Article

Check for updates

Which Robust Regression Technique Is Appropriate Under Violated Assumptions? A Simulation Study

Jaejin Kim[1], Johnson Ching-Hong Li[1]

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

Methodology, 2023, Vol. 19(4), 323–347, https://doi.org/10.5964/meth.8285

Received: 2022-02-02 • **Accepted:** 2023-09-18 • **Published (VoR):** 2023-12-22

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

Corresponding Author: Johnson Ching-Hong Li, Department of Psychology, University of Manitoba, Winnipeg, R3T 2N2, MB, Canada, Phone: 001-204-318-2923. E-mail: Johnson.Li@umanitoba.ca

Abstract

Ordinary least squares (OLS) regression is widely employed for statistical prediction and theoretical explanation in psychology studies. However, OLS regression has a critical drawback: it becomes less accurate in the presence of outliers and non-random error distribution. Several robust regression methods have been proposed as alternatives. However, each robust regression has its own strengths and limitations. Consequently, researchers are often at a loss as to which robust regression method to use for their studies. This study uses a Monte Carlo experiment to compare different types of robust regression methods with OLS regression based on relative efficiency (RE), bias, root mean squared error (RMSE), Type 1 error, power, coverage probability of the 95% confidence intervals (CIs), and the width of the CIs. The results show that, with sufficient samples per predictor (n = 100), the robust regression methods are as efficient as OLS regression. When errors follow non-normal distributions, i.e., mixed-normal, symmetric and heavy-tailed (SH), asymmetric and relatively light-tailed (AL), asymmetric and heavy-tailed (AH), and heteroscedastic, the robust method (GM-estimation) seems to consistently outperform OLS regression.

Keywords

robust regression, OLS regression, outliers, Type I error, power

Among conventional statistical methods, ordinary least squares (OLS) regression is one of the most widely employed statistical analyses used by researchers for prediction and theoretical explanation (Erceg-Hurn & Mirosevich, 2008). Due to its straight-forward interpretation (e.g., linear relationships), easy calculation, and popularity, OLS regression is widely used in many areas of research in fields such as biology, business, education, computer science, psychology, and more (Anderson & Schumacker, 2003; Haupt et al., 2014; Sauvageau & Kumral, 2015; Yellowlees et al., 2016). OLS regression is used to predict dependent variables based on explanatory variables plus errors, which can be presented as

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \tag{1}$$

where y_i denotes the dependent or response variable from $i = 1,\ldots, n$ observations, $x_{ij} = x_{i1}, \ldots, x_{ip}$ (where $j = 1,\ldots, p$) denote p numbers of predictors or explanatory variables, $\beta_0, \ldots, \beta_p$ are the $p + 1$ regression coefficients, and ε_i represents the difference between the actual observed score and the score predicted from a statistical model. The purpose of OLS regression is to minimize the sum of squares of the difference between predicted values and actual observed values. That can be written as

$$\sum \left(y_i - \widehat{y}_i\right)^2 \ = \ \sum_{i=1}^{n} \left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)\right]^2. \tag{2}$$

With the OLS regression model, researchers in psychology can examine whether predictors can significantly predict an outcome measure, which is a highly appealing and widely-employed statistical procedure. For example, in one study (McCrone et al., 2005), OLS regression was used to predict costs based on family history of psychiatric illness. In another study, Mossakowski (2011) used OLS regression to analyze if unfulfilled expectations predict subsequent symptoms of depression.

To use OLS regression properly, researchers need to check whether or not four critical assumptions have been met (Anderson & Schumacker, 2003; Field & Wilcox, 2017; Greene, 2003; Yellowlees et al., 2016). First, errors must follow a normal (N) distribution. Second, explanatory variables should not be correlated with error distribution. Third, homoscedasticity must be achieved; that is, residuals at each level of the predictor variable should have a common variance (Anderson & Schumacker, 2003; Greene, 2003; Yellowlees et al., 2016). Finally, the relationship between the response variable and the explanatory variable should be linear (Field & Wilcox, 2017). When all of these assumptions are met, OLS regression will be the maximal, unbiased linear estimator of the regression coefficients in the population (Field & Wilcox, 2017). However, in reality, this is not what researchers commonly face (Erceg-Hurn & Mirosevich, 2008).

In many research scenarios involving behavioral and social data, these assumptions are violated. One of the assumptions that is often violated is the assumption of homoscedasticity. Erceg-Hurn and Mirosevich (2008) claimed that the presence of heteroscedasticity (HE) is common in real data. In addition to that, normality assumptions are rarely met in practice. Micceri (1989) found that, of 440 large-sample measures related to psychology such as achievement and other common psychometric measures, none

PsychOpen GOLD

were normally distributed among the data they investigated. Approximately 15.2% of 440 distributions followed Gaussian distribution, and most of the distributions followed either a heavy or skewed distribution.

Using OLS regression under violations of assumptions gives rise to several critical problems for researchers. First, when the normality assumption is not met, OLS regression produces lower power and wider confidence intervals. In other words, it weakens the generalizability of data. Moreover, it may inflate the possibility of making a Type 1 error (Anderson & Schumacker, 2003; Erceg-Hurn & Mirosevich, 2008). Even when the normality assumption is not violated, the problems mentioned above can occur with the presence of heteroscedasticity (Brossart et al., 2011). As a result, OLS regression becomes inefficient and generates unstable results when underlying assumptions are not met, which is quite common in practice.

Just as violated assumptions are commonly encountered in reality, so are outliers. As can be seen in Equation 2, each observation is weighted equally by the OLS estimator, and hence, a large difference between an outlier and the mean of all other scores will have a substantial impact in the accuracy of the slope estimates. For this reason, the presence of outliers in data also markedly distorts the efficiency of OLS regression, resulting in erroneous results. Outliers can arise from different sources: man-made or random (Osborne & Overbay, 2004). That is, outliers can be caused by researchers during data entry, incorrect distribution assumptions, and sampling error, or by random chance when collecting samples from a population.

Outliers have different influences on the estimation of regression coefficients, depending on their location. The influence of an outlier may be much more severe when it lies on the x-axis, called the leverage point, than when it lies on the y-axis (Anderson & Schumacker, 2003). The leverage point can be either good or bad (Rousseeuw & Leroy, 2003). A good leverage point, which is away from the bulk of the points but close to a regression line, reduces the standard error. However, when the location of a data point is far away from the rest of the data points and from the line of best fit, called bad leverage, it can pull the regression line towards the outlier's location. Therefore, outliers in data not only indicate the assumptions that normality and homogeneity may not hold, but also seriously impact the result based on OLS regression including inaccurate intervals, lowering statistical power, and Type 1 and Type 2 errors (Brossart et al., 2011).

One of the common approaches to handling these outliers is to transform data using logarithms or square roots (Grissom, 2000). Transformation may be a valuable option; however, it often gives rise to more problems. Transforming data often fails to restore normality and homoscedasticity and is not adequate to deal with outliers (Wilcox, 2022). Moreover, it also leads researchers to interpret data in an inaccurate way by changing the original construct (Osborne, 2003).

The other option that researchers commonly use is to discard outliers. Yellowlees et al. (2016) argued that this method, however, can only be used if an outlier can be traced back to an error that occurred during an experiment (e.g., wrong dose of a drug or product). Finney (2009, p. 51) also stated: “the danger [is] that scientists bias their conclusions by removing data that deviate markedly from current ideas of truth.” They warned: “[n]ever discard an apparent outlier unless there is strong evidence that it was the product of a measurement or other form of observation that suffered a gross mistake or accident, this misfortune being unrelated to any experimental treatment under investigation.” Hence, this method may only be advisable when researchers clearly know the cause of outliers (Yellowlees et al., 2016). Unfortunately, it is hard for researchers to objectively define outliers especially when the data have many explanatory variables (Maronna et al., 2006). Moreover, the presence of outliers sometimes disguises other outliers (called a masking effect; Wilcox, 2022). On the other hand, retaining outliers could still result in a highly misleading result, if the goal of a research study is to characterize the nature of the association among the bulk of participants.

Robust regression can be an alternative method to deal with outliers and assumption violations. Despite the potential of robust regression, there is no single study that comprehensively integrates these methods and systematically evaluates their accuracy in a Monte Carlo experiment. The current research fills this research gap. The following section discusses the mathematical and computational details of these methods.

Robust Regression

Robust regression is a modern method conceptualized many decades ago. In the 1950s, Siegel (1956) stated that non-parametric and robust techniques of hypothesis testing are best suited to behavioral sciences data. However, robust regression has only recently been studied due to advancements in computer technology (Anderson & Schumacker, 2003). Unlike OLS regression, which gives weights to outliers, robust regression reduces the impact of the outliers by weighing them down. This allows researchers to take outliers into account in the statistical model rather than using other, potentially problematic methods to deal with them.

Two pivotal concepts need to be addressed to understand robust regression methods: breakdown point, and relative efficiency (Anderson & Schumacker, 2003). The breakdown point measures the minimum proportion of points that are needed to make a statistical estimate, such as regression slope, arbitrarily large or small. The breakdown points vary between 0, or $1/n$ and 50%, or $n/2$. In other words, an estimator with a 0% breakdown point does not efficiently prevent the regression equation from being influenced by regression outliers or bad leverage points. OLS regression has the breakdown point of 0%; consequently, the presence of one outlier or bad leverage point can render the data inefficient. On the other hand, robust regression estimators with a 50% breakdown point can contain as many as 50% bad leverage points without making a statistical estimate being arbitrarily large or small. It is noteworthy that even though replacing the OLS regression with robust estimators having a high breakdown point appears to be a reasonable solution when data contains outliers or bad leverage points, some recent studies (e.g., Wilcox, 2022; Wilcox & Xu, 2023) have showed that the presence of a few outliers or bad leverage points could still have a noticeable impact on the slope estimates. Or, stated differently, robust estimators can only ensure that having a few outliers or bad leverage points will not lead to an arbitrarily large or small statistical estimate, but they do not guarantee that the slope estimates are not substantially influenced by outliers or bad leverage points, thereby leading to a misleading conclusion of the true association in practice.

Another key concept is relative efficiency, which refers to the extent to which robust regression performs like OLS regression when error distribution follows a normal pattern. The relative efficiency is determined by dividing OLS regression mean square error (MSE) by the robust regression MSE, which can be expressed as (Anderson & Schumacker, 2003)

$$\text{Relative efficiency} = \text{RE} = \frac{MSE\ OLS}{MSE\ Robust}$$

This is often expressed using a percentage value ranging between 0% and 100% or more. For example, if one robust regression technique has 75% relative efficiency, this means that the method is 75% as efficient as OLS regression. While robust regression may suffer from a slightly lower efficiency than OLS regression when the normality and homoscedasticity assumptions are met, robust regression is expected to produce much more accurate (or robust) estimates and results if the assumption is violated, which is crucial for obtaining more accurate statistical results in psychological research.

Types of Robust Regressions

Least-Square-Fit-CI (lsfitci) Approach

One approach to dealing with violation of the normal-error assumption is bootstrapping residuals (ε_i) instead of raw scores (x and y) to generate empirical distribution and standard error of residuals for the construction of $1 - \alpha$ CI, where α is the type 1 error rate. By locating the $\alpha/2$ and $1 - \alpha/2$ percentiles of the bootstrap residuals, it is expected that the CI could be asymmetrical, thereby adjusting for any non-normal residuals. This method is known as “least square fit CI” (lsfitci) in Wilcox (2022), and its performance is evaluated in this study.

Heteroscedastic-Consistent (HC) Standard-Error Approaches

HC standard-error approaches (Huber, 1967; Long & Ervin, 2000; White, 1980) can be used to fit a regression model that contains heteroscedastic errors. Among them, HC3 method was developed for studies with large sample sizes (Wilcox, 2022). This method can effectively replace the bootstrap standard error estimate by

HC3 = $S = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag\left[r_i^2/\left(1-h_{ij}\right)^2\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, where r_i is the residual for i = 1,..,n, $h_{ij} = \mathbf{x}_\mathrm{i}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathrm{i}'$, and

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

where $\mathbf{x}_\mathrm{i}$ is the ith row of $\mathbf{X}_i$. Consequently, the 1 − α CI surrounding β_p is $b_j \pm tS_j$, where t is the $1-\alpha/2$ quantile of t distribution with $n-p-1$ degrees of freedom. Another approach is called HC4, which is a modified version of HC3 that was found to be better for more general use. That is, HC4 = $S = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'diag\left[r_i^2/\left(1-h_{ij}\right)^{d_{ii}}\right]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$.

HC-Robust Wild Bootstrap (W-B) Approach

The W-B approach was originally developed by Wu (1986), and it can be used to produce unbiased estimates of regression models with heteroscedastic errors. The W-B approach resamples the multiway, clustered heteroscedastic error terms to estimate the bootstrap dependent variable scores for constructing the CI surrounding the slope parameters. Roodman et al. (2019) have advanced and modified Wu's (1986) approach by developing a fast W-B approach that can efficiently calculate bootstrap test statistics and implements a HC-robust W-B procedure for constructing the CI via their developed R package (fwildclusterboot; Fischer et al., 2023), and the performance of this approach is evaluated in this study.

Least Median of Squares (LMS) Estimator

The LMS estimator was developed by Rousseeuw (1984). Unlike OLS regression, using the sum of the squared residuals, the LMS-estimator uses the median of the squared residuals. This can be expressed as

$$\min M\left[y_i - \left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)\right]^2 = \min M(r_i^2), \tag{3}$$

where M is the median. The LMS-estimator is the first method that achieves the breakdown point 0.5; therefore, it is resistant to outliers. The LMS function in R does not provide an analytic method for the standard error, but it can be estimated through bootstrapping that locates the $\alpha/2$ and $1-\alpha/2$ percentile bootstrap slopes based on the LMS-estimator for the 1 − α CI (called LMS-B in this study). However, it has a critical limitation: the relative efficiency of the LMS-estimator to OLS regression is 0 due to $n^{-1/3}$ convergence. For this reason, the LMS-estimator is not practically useful, but it plays a significant role in other robust methods such as the MM-estimator, which is described below (Andersen, 2008).

Least Trimmed Squares (LTS) Estimator

Another robust estimator developed by Rousseeuw (1984) is called LTS, which is defined as

$$\text{Minimize} \sum_{i=1}^{h} \left(r_i^2\right), \tag{4}$$

where $r_{(1)}^2 \leq \ldots \leq r_{(h)}^2$ are the ordered squared residuals following ascending order. The breakdown point of 0.5 can be achieved with $h = [n/2] + [(p+1)/2]$, which is commonly used (Wilcox, 2022). Although the LTS-estimator may have a high breakdown point depending on h, its efficiency is very low, about 8% (Stromberg et al., 2000). Nevertheless, this method has some value insofar as it is used as an initial estimate for other robust methods (Andersen, 2008). Another related approach is the use of bootstrapping that locates the $\alpha/2$ and $1 - \alpha/2$ percentile bootstrap slopes based on the LTS-estimator for the $1 - \alpha$ CI, which is labelled as LTS-B.

Maximum Likelihood Type Estimation (M-estimator)

The M-estimator proposed by Huber (1973) minimizes

$$\sum_{i=1}^{n} \psi\left(r_i\right), \tag{5}$$

where ψ is a robust loss function with a unique minimum at zero. The robustness of the M-estimator depends on a robust loss function that researchers choose. One commonly used function is Huber's p function, expressed as

$$\psi(\mathrm{x}) = \left\{\begin{matrix} \frac{1}{2}x^2 & \text{for } |x| < c \\ c|x| - \frac{1}{2}c^2 & \text{for } |x| \geq c \end{matrix}\right\}, \tag{6}$$

where c is a tuning constant which can be adjusted to control asymptotic efficiency. When outliers lie in the y-axis, the Huber M-estimator, in general, is more efficient than OLS regression against outliers. However, it does not consider a leverage point. If there is an outlier in the x-axis, the Huber's M-estimator is not better than OLS regression; therefore, the breakdown point is $1/n$ (Wilcox, 2022).

Generalized M-Estimator (GM-Estimator)

Due to the limitation of the M-estimator which does not consider leverage points, a generalized M-estimator was developed to guard against leverage points by adding some weight, ω_i, to x_i values. Mallow (1973, as cited in Krasker & Welsch, 1982, p. 596) proposed the GM-estimator, which can be expressed as

$$\sum_{i=1}^{n} \omega_i \psi\left(r_i / \widehat{\tau}_i\right) x_{ij} = 0 \tag{7}$$

where $x_{io} = 1$, and $j = 0, \ldots, p$. Mallow used $\omega_i = \sqrt{1 - h_{ii}}$ based on a condition if $h_{ii} > h_{jj},\ \omega_i < \omega_j$. That is, high leverage points to x_i receive less weight than low leverage points to x_i. As a result, this method gives less weight to good leverage points resulting in a loss of efficiency (Andersen, 2008).

To solve the low efficiency issue by using Mallow's weight, Schweppe proposed a different solution (Handschin et al., 1975), expressed as

$$\sum_{i=1}^{n} \omega_i \psi\left(r_i / \left(\omega_i \widehat{\tau}_i\right)\right) x_{ij} = 0, \tag{8}$$

where $j = 0, \ldots, p$. The idea of Equation 8 is to use different weight values according to the size of the residuals. In other words, Schweppe tried to solve the limitation of Mallow's weight by dividing r_i by ω_i. Even though Schweppe's estimator may provide a better option for dealing with leverage points than the regular M-estimator, its break point is less than 0.5 (Maronna et al., 2006), and especially low with a large number of predictors (Andersen, 2008).

Schweppe One-Step (S1S) Estimator

Coakley and Hettmansperger (1993) expanded from the Schweppe's estimator and developed the S1S estimator, expressed as

$$\widehat{\beta} = \widehat{\beta}_0 + \left[\sum_{i=1}^{n} \psi'\left(\frac{r_i}{\omega_i}\right) x_i x_i'\right]^{-1} \times \sum_{i=1}^{n} \omega_i \psi\left(\frac{r_i}{\omega_i}\right) x_i \tag{9}$$

where ω_i is determined by using the same criterion that the original Schweppe's estimator used. The S1S-estimator is different from the two GM-estimators mentioned above in that it can achieve 95% efficiency when the error term is normally distributed. Moreover, it achieves a breakdown point of 0.5 by using the LTS-estimator as an initial estimator (Andersen, 2008). Wilcox (2022) states that the S1S-estimator can be effective when the sample size is large, and ε follows a normal distribution. However, it becomes inefficient when samples sizes get smaller.

MM-Estimator

Another popular robust technique derived by Yohai (1987) is MM-estimator, so called because it calculates the final estimates by employing more than one M-estimation. Three steps are involved to find the MM-estimator. The first step is to compute initial estimates of the coefficients $\widehat{\beta}$ with high breakdown points, 0.5, using s-estimation.

Then, a robust M-estimate of scale $\widehat{\sigma}$ of the residuals is calculated by using the initial estimation (Maronna et al., 2006). The robust scale of $\widehat{\sigma}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\psi\left(\frac{r_i}{\widehat{\sigma}}\right) = 0.5 \tag{10}$$

The final step is to compute the regression parameters by solving the following equation (Wilcox, 2022):

$$\sum_{i=1}^{n}\psi(r_i/\widehat{\sigma})x_{ij} = 0, \tag{11}$$

where $j = 0, \ldots, p$, and ψ, is Tukey's biweight which is commonly used. Tukey's biweight is expressed as

$$\psi(x) = \left\{\begin{matrix} x(1 - x^2)^2 & |x| < 1 \\ 0 & |x| \geq 1 \end{matrix}\right\} \tag{12}$$

where $\widehat{\sigma}$ in Equation 10 is a robust M-estimate of scale, and Tukey's biweight is used as a redescending function in Equation 10.

$$\psi(r_i; c) = \left\{\begin{matrix} \frac{r_i}{\widehat{\sigma}}\left(\left(\frac{r_i}{c\,\widehat{\sigma}}\right)^2 - 1\right)^2 & |r_i/\widehat{\sigma}| \leq c \\ 0 & |r_i/\widehat{\sigma}| \geq c \end{matrix}\right\}. \tag{13}$$

When c is 4.685, the relative efficiency of the MM-estimator is 95% to OLS regression. Under normality, it has a high breakdown point, 0.5, and relative efficiency, 95%, to OLS regression (Wilcox, 2022). Another related approach is the use of bootstrapping that locates the $\alpha/2$ and $1 - \alpha/2$ percentile bootstrap slopes based on MM-estimator for the $1 - \alpha$ CI; this method is called MM-B in this study.

S-Estimators

Rousseeuw and Yohai (1984) proposed the S-estimators that estimate slope and intercept values with the goal to minimize some measure of scale corresponding to the residuals. Indeed, the conventional, least squares approach is regarded as one type of S-estimator that minimizes the variance of the residuals. In this case, replacing the variance with some measure of location that is robust to outliners is the goal of using S-estimators for estimating robust slopes and intercepts. Wilcox (2022) mentioned that S-estimators may have some practical value, but no study has examined their empirical performance via simulation studies. One approach is the Nelder-Mead method (SNM; e.g., Olsson & Nelson, 1975). According to Wilcox (2022); let $R_i = y_i - b_i x_{1i} - \ldots b_p x_{ip}$, and the SNM approach searches the values of $b_i, \ldots, b_p$ such that the standard error estimate (S) is minimized through some measure of scale based on the values of $R_i \ldots, R_n$. Consequently, the intercept is $b_0 = M_y - b_1M_1 - \ldots b_pM_p$, where M_y and M_j are the medians of the y values and of the x_{ij} scores, where $i = 1, \ldots, n$.

E-Type Skipped Estimator

The purpose of using an E-Type (or error-type) skipped estimator is to remove or decrease the influence of any outliers existing in a dataset on fitting a regression model. This estimator often begins by running preliminary fit to search for any outlier residuals, removing or downweighing those outliers, and fitting a regression model based on the remaining data. Wilcox (2022) supposes that M_r is the median of the residuals, and MAD_{r} (median absolute deviation) is the median of the values $|r_i - M_r|, \ldots, |r_n - M_r|$. Consequently, for an ith point (x_i, y_i) with $|r_i - M_r| > 2(MAD_r)/.6745$, this point is deemed an outlier. The slope and intercept estimates will then be based on the data points that are not declared as outliers.

Methods Based on Robust Covariances

Replacing conventional covariances with robust covariances in fitting a regression model is a general approach that leads to robust intercept and slope parameter estimators of the model (ROB; Huber, 1981). In its simplest form with one predictor, the slope of the OLS regression line is $\beta_1 = \sigma_{xy}/\sigma_x^2$, in which the numerator can be replaced by a robust covariance estimate between x and y, and the denominator can be replaced by a robust variance estimate of x. One common approach for estimating robust variance and covariance is the biweight midcovariance that estimates the variability and co-variability of the scores based on the robust location (medians) and robust distance (MAD) that exist in a dataset.

Quantile (QUA) Regression

Another robust approach is estimating regression parameters based on minimizing the summation of the absolute of the residual scores, $\sum |r_i|$. According to Koenker and Bassett (1978), researchers could estimate the qth quantile of y scores given x scores. Suppose $\rho_q(u) = u(q - I_{u<0})$, where I is the indicator function. Hence, the regression model is estimated by minimizing $\sum \rho_q(r_i)$. When $q = 0.5$ (or 50th quantile), it refers to the least absolute value of the estimator which, in turn, leads to an estimate of the median of y scores, a robust measure of location, given x scores.

In sum, it has been suggested that robust regression methods outperform OLS regression when outliers exist in the data (Andersen, 2008; Anderson & Schumacker, 2003; Brossart et al., 2011; Finger, 2010; Maronna et al., 2006; Mercer et al., 2015; Sauvageau & Kumral, 2015; Wilcox, 2022; Wilcox & Keselman, 2004; Yellowlees et al., 2016). However, no robust regression technique is universally superior, because each regression method has strengths and limitations. Depending on the situation, one may be more appropriate than another. In terms of handling leverage points, the S1S-estimator may be the best option. However, the S1S-estimator becomes less efficient with a small sample size, in which case the MM-estimator may be more appropriate (Andersen, 2008). In general, MM-type robust estimation may be the preferred choice in terms of relative efficiency, bias, and testing the null hypothesis (Anderson & Schumacker, 2003). When outliers are located in the y-axis, the Huber M-estimator would be appropriate as well, because it has a .5 breakdown point regarding outliers in the y-axis (Yellowlees et al., 2016). Therefore, the purpose of this research study is to compare robust regression methods using different settings to provide other researchers with information that will help them select the most appropriate robust regression methods when errors violate normality and homoscedasticity assumptions in psychological research.

Method: A Monte Carlo Simulation

To provide a better understanding of the selection of appropriate robust regression methods, we used a Monte Carlo simulation study to compare OLS and robust regression methods under a variety of conditions that researchers commonly face. The simulation was conducted based on multiple regression models with two independent predictors that can be expressed as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i \tag{14}$$

To represent varied research conditions, we included variations across the following research variables: sample size per predictor, slope, and different types of error distribution.

Sample Size Per Predictor

Several rules have been proposed for minimum required sample ratio per predictor to conduct multiple regression analysis (Miller & Kunce, 1973; Schmidt, 1971). Schmidt (1971) recommends 15 to 20 samples per predictor, whereas Miller and Kunce (1973) argue that there should be 30 samples per predictor for accurate regression analysis. In response, and to provide more representative results, we added two more sample size variables, thus including 20:1, 30:1, 50:1, and 100:1 (*n*; sample size per predictor) in this study.

Slope

For slope, we selected values of 0, .3606, and .5099 (Cohen, 1988), which correspond to the zero effect, medium effect ($.3606^2$ = 13% of the variance of *y* explained by *x*), and substantial effect ($.5099^2$ = 26% of the variance of y explained by x), to create nine combinations: (0, 0); (0, .3606); (0, .5099); (.3606, 0); (.3606, .3606); (.3606, .5099); (.5099, 0); (.5099, .3606); and (.5099, .5099).

Error Distribution

For error distribution, we adopted the same criteria used by Yuan and MacKinnon (2014) and Wilcox (2022). First, a normal distribution, $N \sim (0, 1^2)$, was simulated. Second, following Yuan and Mackinon, a mixed normal distribution (MN) with 90% random errors was generated from $N \sim (0, 1^2)$ and 10% random errors generated from $N \sim (0, 10^2)$. It is noteworthy that MN is a type of symmetric distributions with heavy tails on both ends, and it has been widely employed and tested in previous simulation studies (e.g., Algina et al., 2005). Third, a heteroscedastic error term, in which the variance used to generate the random error score of each simulated participant depends on his or her x_{ip}, $N \sim (0, x_{ip}^2)$, where $i = 1,..,n$, and $p = 1$ or 2, was simulated (Yuan & MacKinnon, 2014). The remaining three distributions followed Wilcox's (2022) method for simulating asymmetric and/or heavy tailed error distribution for testing the performance of robust regressions. That is, the fourth distribution was a symmetric and heavy-tailed (SH) distribution based on a g-and-h distribution with (g, h) = (0, 0.5). The fifth distribution was an asymmetric and light-tailed (AL) distribution with (g, h) = (0.5, 0). The sixth distribution was an asymmetric and heavy-tailed (AH) distribution with (g, h) = (0.5, 0.5).

In summary, four sample sizes per predictor, nine combinations of slopes, and six types of error distributions were evaluated. This factorial design created a total of $4 \times 9 \times 6 = 216$ different conditions. Each of the 216 conditions were replicated 1,000 times. For bootstrapping, the simulated x and y scores were resampled 1,000 times for constructing the 95% bootstrap percentile intervals. In sum, this design produced a total of $216 \times 1,000 \times 1,000 = 216{,}000{,}000$ simulated data sets for evaluation. OLS-regression estimates and CI, nine robust regression estimates and analytic-based CIs (i.e., LTS-estimator, M-estimator, GM-estimator, S1S-estimator, MM-estimator, S-estimator, E-type, ROB, and QUA regressions), one robust, analytic-based CI for OLS-regression estimates (i.e., lsfitci), and five bootstrap-based CIs (i.e., HC3, HC4, LMS-B, LTS-B, and MM-B) described above were performed on these simulated data sets in order to compare the accuracy of their results. We used the statistical software, R, to conduct our simulation (R Core Team, 2023), and the code is shown in the Supplementary Materials.

Criteria

The criteria we used to compare the regression methods were relative efficacy, bias, RMSE, Type I error, power, coverage probability of the 95% CI, and width of the CI. For relative efficacy, higher percentages are desirable. For example, if the relative efficacy of a robust approach is .98, this means that it can maintain 98% of efficiency compared to the conventional OLS estimates, when the normality and homoscedasticity assumptions are met. Regarding bias, when a regression method generates a slope farther away from the true slope, this is considered bias. Consequently, bias, in this study, is defined as the difference between the mean of the 1,000 replicated slopes minus the true slope (i.e., $bias = \bar{b} - \beta$, where $\bar{b}$ is the mean of 1,000 replicated observed slopes in each simulated condition, and β is the true slope value manipulated in the condition). Evaluating the bias of the slope estimates is insufficient because it does not measure the variability of those values. The root mean square error (RMSE) and variance of the 1,000 replicated slope values are also included. RMSE is defined by $\sqrt{\sum_{r=1}^{1,000}(b_r - \beta)^2/1,000}$, which measures the average (squared) distance between each of the 1,000 replicated slope values with the true value. Confidence width is used to evaluate the precision and sampling error of the slope estimates. A narrower confidence width indicates a more precise estimate. For Type 1 error, we set its error rate at $\alpha = .05$ level (two-tailed test), which is commonly used as a criterion in psychological research. When the true slope was set at 0, we examined the number of times (or probability) that a regression method would lead to an incorrect decision (i.e., rejecting the null hypothesis: $\beta = 0$) out of 1,000 replications for each of the 216 manipulated conditions. By the same token, when the true slope was set at .3606 and .5099, we evaluated the number of times (or probability) that a regression method would lead to a correct decision (i.e., rejecting the null hypothesis: $\beta = 0$) out of 1,000 replications for each of the 216 manipulated conditions. The coverage probability examines the probability a 95% CI has spanned a true slope parameter value. Theoretically speaking, of the 1,000 replications, the number of the 95% CIs that has spanned the true parameter value is expected to be 950 (or coverage probability = 95%). In practice, sampling error exists, and hence, an observed coverage probability ranging from .925 to .975 yielded by a regression method is deemed desirable (Chan & Chan, 2004).

Results

Relative Efficiency (RE)

The results of RE of each robust regression method compared to OLS regression, with normal error distribution, are presented in Table 1[1] (for all tables, see Supplementary Materials). When errors were normally distributed, three types of RE results were observed. The first type was observed by LMS and LTS, which produced the least efficient RE with a range from .761 to .778 when the sample size was small (20). The second type was observed by S, E, and QUA, and they were regarded as moderate RE, which ranged from

[1] This table and all tables subsequently referenced throughout the article can be found in the Supplementary Materials.

.935 to .960 with a small sample size of 20. The third type was found in the M, GM, S1S, MM, and ROB approaches leading to the highest RE, which ranged from .958 to .996 with a small sample size of 20. Comparing the effects of different manipulated factors, the sample size was found to be the most influential. As the sample size got larger, the discrepancy between the efficiency of robust regression and OLS regression got smaller. When the sample size was larger (e.g., *n* = 100), most of the robust regression methods (except LMS and LTS) were nearly the same in efficiency, ranging from 0.982 to 0.999. Therefore, this result indicates that there is no obvious or substantial loss of efficiency based on the robust regression methods compared to OLS regression, especially when *n* is greater than 100.

Bias, RMSE, and Variance of the Slope Estimates

Given that the patterns of the results are similar between Slopes 1 and 2, the following paragraphs focus on the results based on Slope 1 estimates only. Table 2 shows that, when errors were normally distributed, all the slope values were close to the true slope, with biases ranging from -.020 to .064 with a mean of .005 and median of .001. That is, no regression method showed bias in slopes with normally distributed errors. When errors were symmetric with long tails (MN and SH; Table 3 & 4), the biases were still appropriate, and they ranged from -.054 to .08 with a mean of .005 and median of .001. When errors were lightly skewed (AL; Table 5), all the biases were reasonable, range = (-.011, .055), mean = .003, median = .000; however, when errors were heavily skewed (AH; Table 6), the range of the bases became large based on OLS, range = (-.112, .105), mean = .003, median = -.001, and the other robust methods produced similar patterns of biases, range = (-.023, .080), mean = .006, median = .001, as in other error distributions. When errors were heteroscedastic (Table 7), first, the MM method consistently produced an error message in R, and hence, it was inappropriate for evaluation. Second, the OLS method resulted in reasonable biases, range = (-.028, .031), mean = .001, median = -.001, as in most other robust methods, range = (-.026, .025), mean = .000, median = .000, except for the S-estimator that resulted in biases ranging from .065 to .144 with a mean of .102 and median of .101.

When errors were normal (Table 8), the OLS method resulted in the smallest RMSE values, range = (.068. .258), mean = .140, median = .117. This is reasonable as OLS should be the most precise in estimating the slope values when the normality assumption is met. All the robust methods resulted in larger RMSE values, range = (.070, .613), mean = .185, median = .141. When errors were symmetrical with longer tails (i.e., MN and SH; Tables 9 & 10), OLS produced larger RMSE values, range = (.230, 1.715), mean = .570, median = .449, than all the robust methods, range = (.076, .663), mean = .209, median = .167. When errors were skewed, the RMSE of OLS depended upon whether the tail was light or heavy. With AL (Table 11), range = (.086, .324), mean = .170, and median = .141. Some of the robust methods had even larger RMSE than OLS, e.g., LMS, range =

(.142, .569), mean = .288, median = .234; S, range = (.087, .334), mean = .177, median = .144; E, range = (.082, .335), mean = .174, median = .142; and QUA, range = (.086, .332), mean = .175, median = .144, whereas the remaining robust approaches resulted in smaller RMSE values, range = (.067, .331), mean = .152, median = .126. With AH (Table 12), RMSE became noticeable larger for OLS, range = (.545, 4.198), mean = 1.344, median = 1.075, indicating a larger bias. All the robust methods produced smaller RMSE, range = (.082, .649), mean = .214, median = .173). When errors were heteroscedastic (Table 13), OLS also resulted in larger RMSE values, range = (.260, .832), mean = .473, median = .419, than all the robust approaches, range = (.000, .510), mean = .230, median = .190.

The patterns of the variance of 1,000 replicated slope estimates across 216 conditions are identical to the patterns of the RMSE values, as they measured the variability of those estimates. That is, the variance was found to be smaller for OLS, range = (.005, .030), mean = .015, median = .014, than other robust methods, range = (.005, .167), mean = .029, median = .020, when errors were normal (Table 14). When errors were symmetrical with long tails (i.e., mixed-normal and SH; Table 15 & 16), the variance was larger for OLS, i.e., range = (.05, 5.385), mean = .406, median = .255, than the robust methods, i.e., range = (.006, .168), mean = .034, median = .028. When errors were AL (Table 17), the variance of the OLS estimates, range = (.007, .043), mean = .022, median = .020, was similar to most of the robust methods, range = (.004, .047), mean = .019, median = .017, except LMS, range = (.021, .121), mean = .060, median = .053, suggesting that the variance of the OLS slopes was less influenced by an asymmetric, light tail. On the contrary, the variance of the OLS slopes, i.e., range = (.384, 24.891), mean = 2.464, median = 1.055, became noticeably larger than the robust methods, i.e., range = (.007, .134), mean = .033, median = .030, with an asymmetric, heavy tail (Table 18). When errors were heteroscedastic (Table 19), the variance of OLS slope estimates, i.e., range = (.068, .353), mean = .193, median = .178, was larger than the robust methods, i.e., range = (.0009, .112), mean = .040, median = .036.

Type I Error

When errors were normal (Table 20), the Type 1 error rates were well protected for the OLS method, range = .044, .062, mean = median = .051. Some robust methods (e.g., HC3, HC4, W-B, MM-B, and ROB) produced Type 1 error rates similar to those obtained by OLS. Of the remaining robust methods, some resulted in slightly smaller (or more conservative) Type 1 error rates, e.g., LMS-B: range (.000, .007), mean = median = .003; LTS-B: range (.004, .036), mean = .015, median = .012; GM: range (.031, .060), mean = median = .043; S1S: range (.013, .064), mean = .032, median = .003; S: range (.010, .033), mean = median = .021; E: range (.007, .039), mean = .021, median = .020; QUA: range (.018, .042), mean = .028, median = .026, while others produced higher Type 1 error rates, e.g., lsfitci range (.056, .095), mean = .074, median = .073; LTS: range (.115, .163), mean = .146, median = .148; M: range (.036, .082), mean = .059, median = .061; MM: range (.,035, .080), mean = .058, median = .059.

When errors were mixed-normal, SH, AL, and AH (Tables 21–24), the Type I error rates produced by OLS were surprisingly comparable with those observed in the normal distribution: range (.038, .065), mean = .051, median = .050. Four robust methods, W-B, range (.031, .061), mean = .047, median = .047; GM, range (.029, .059), mean = .041, median = .040; MM-B, range (.032, .065), mean = .047, median = .046; and ROB, range (.033, .067), mean = .046, median = .045, yielded reasonable Type 1 error rates close to the nominal level of .05. Nine robust methods, HC3, HC4, LMS-B, LTS, LTS-B, S1S, S, E, and QUA, produced Type I error rates slightly smaller than the nominal level: range (.000, .066), mean = median = .023. Three robust methods, LTS, N, and MM, yielded Type I error rates higher than the nominal level of .05, range (.031, .135), mean = .072, median = .062.

When errors were heteroscedastic (Table 25), OLS performed poorly with less protected (or higher) Type I error rates: range (.361, .409), mean = median = .378. Indeed, only four robust methods, HC3, range (.043, .065), mean = .056, median = .060; HC4, range (.041, .057), mean = .051, median = .054; W-B, range (.047, .068), mean = .055, median = .053; and GM, range (.037, .058), mean = .048, median = .047, produced reasonable Type I error rates. Eight robust methods, LMS-B, LTS-B, S1S, MM-B, S, E, ROB and QUA, yielded smaller Type I error rates: range (.004, .059), mean = .027, median = .028. The remaining three robust methods, lsfitci, LTB, M, and MM resulted in larger Type I error rates: range (.072, .606), mean = .390, median = .505.

Power

When errors were normal (Table 26), the power rates produced by OLS were the largest: range (.566, 1), mean = .882, median = .949. Of the remaining robust methods, some of them, including lsfitci, HC3, HC4, W-B, LTS, M, and MM, yielded power rates similar to those obtained by OLS: range = (.512, 1), mean = .871, median = .935. GM, MM-B, and ROB produced power rates slightly smaller than OLS: range (.403, 1), mean = .816, median = .878. The rest of the robust methods, LMS-B, LST-B, S1S, S, E, and QUA, produced much smaller power rates: range (.019, 1), mean = .585, median = .584. The factors that influenced the power rates were the same for all the approaches. First, when *n* increases, power rates increase. Second, when β_1 increases, power rates increase. Third, there is no obvious relationship between the size of β_2 and power rates.

When errors were mixed normal, SH, AL, and AH (Table 27–30), the OLS produced the smallest power rates, range (.125, 1), mean = .47, median = .329. This pattern of results is reasonable because the standard error should be more biased with the longer tails, and hence, this affects the precision of the estimates and likelihood of detecting significant results. Of the remaining methods, most of them produced power rates larger than those obtained by OLS. The LTS and MM led to the relatively larger power rates in the mean range of .80, i.e., range (.402, 1), mean = .829, median = .874. The M, GM, MM-B, and ROB produced power rates in the mean range of .70, i.e., range (.262, 1), mean = .745, median = .789. The S1S, S, E, and QUA fell in the mean range of .60, i.e., range

(.153, 1), mean = .658, median = .654. The remaining robust methods, lsfitci, HC3, HC4, W-B, LMS-B, LTS-B, resulted in power rates that fell in the mean range of .4, range (.024, 1), mean = .461, median = .399.

When errors were heteroscedastic (Table 31), the OLS produced power rates ranging from .458 to .865, with a mean (or median) of .634 (or .614), although one should be cautious about the use of the OLS because of the highly unprotected Type I error rates. Of the remaining robust methods, LTS, M, S1S resulted in the largest power rates, range (.301, 1), mean = .859, median = .928. The LTS-B, GM, S, E, and QUA resulted in power rates like those observed in OLS, range (.240, .995), mean = .629, median = .607. The other approaches, lsfitci, HC3, HC4, W-B, LMS-B, MM-B, and ROB produced power rates smaller than those obtained by OLS, range (.134, .973), mean = .402, median = .344.

Coverage Probability and Width of CI

When errors were normal (Table 32), the coverage probabilities yielded by the OLS method were desirable: range (.928, .962), mean = median = .946. Of the 36 conditions, 36 (or 100%) produced coverage probabilities within the criteria of (.925, .975). The widths of the CI ranged from .278 to .654 with mean = .460 and median = .457 (Table 33). Of the remaining robust methods, eight of them, HC3, HC4, W-B, M, GM, MM, MM-B, and ROB, led to coverage probabilities comparable to the OLS approach, range (.918, .972), mean = .950, median = .951. Of the 36 conditions, 36 (or 100%) yielded by HC3, HC4, W-B, GM, MM, MM-B, and ROB fell within (.925, .975), whereas 35 (or 97.2%) produced by M fell within (.925, .975). The widths were slightly wider than those obtained by OLS, range (.282, .837), mean = .513, median = .516, and this pattern is reasonable as the OLS-based CI is expected to be the most precise with normal errors. The remaining seven robust methods, lsfitci, LMS-B, LTS-B, S1S, S, E, and QUA, were subpar. LMS-B, LTS-B, S1S, S, E, and QUA produced coverage probabilities larger than expected, range (.936, 1), mean = .981, median = .982. Of the 36 conditions, only around 11.5 conditions (or 31.9%), on average, fell within the criteria of (.925, .975). The widths were wider than those obtained by OLS, range (.322, 2.294), mean = .963, median = .883. On the contrary, lsfitci and LTS resulted in smaller coverage probabilities, range (.814, .932), mean = .879, median = .886. Of the 36 conditions, none of the LTS's (or one of the lsfitci's) conditions fell within the criteria of (.925, .975). The widths were narrower than those observed in OLS, range (.253, .612), mean = .425, median = .420.

When errors were mixed-normal, SH, AL, and AH (Table 34 to 37), the OLS method still produced good coverage probabilities: range (.925, .965), mean = median = .946. Of the 144 conditions, 144 (or 100%) yielded coverage probabilities within (.925, .975). As shown in Table 38–41, the widths ranged from .333 to 3.752 with mean = 1.514 and median = 1.550. Five robust methods, M, GM, MM, MM-B, and ROB, yielded desirable coverage probabilities: range (.919, .981), mean = .951, median = .952. Of the 144 conditions, 142.4 (or 98.9%), on average, led to coverage probabilities within (.925, .975). The widths were noticeably narrower than those found in OLS, range (.281, 1.225), mean = .625, median = .581. The W-B method yielded desirable coverage probabilities, range (.936, .969), mean = .952, median = .953, but its widths were noticeably wider, range = (.337, 3.610), mean = 1.401, 1.314. Eight robust methods, HC3, HC4, LMS-B, LTS-B, S1S, S, E and QUA, produced larger coverage probabilities, range (.934, 1), mean = median = .977. Of the 144 conditions 64.1 (or 44.5%) resulted in coverage probabilities within (.925, .975). The widths ranged from .289 to 3.585 with mean = .625, median = .581. Two robust methods (lsfitci and LTS) produced smaller coverage probabilities, range (.838, .943), mean = median = .895. Of the 144 conditions, only 6 (or 4.2%) yielded coverage probabilities within (.925, .975). The widths ranged from .304 to 2.960 with mean = 1.277 and median = 1.298.

When errors were heteroscedastic (Table 42), the OLS produced undesirable coverage probabilities: range = (.584, .654), mean = .615, median = .612. Of the 36 conditions, 0 fell within (.925, .975). As shown in Table 43, the widths ranged from .464 to .989 with mean = .728 median = .735. Six robust methods, HC3, HC4, W-B, GM, E, and QUA, led to desirable coverage probabilities, range (.930, .979), mean = .951, median = .950. Of the 36 conditions, 35.7 (or 99.1%), on average, fell within (.925, .975). It is noteworthy that GM, range (.937, .967), mean = median = .950; HC3, range (.930, .960), mean = .944, median = .943; HC4, range (.937, .965), mean = .949, median = .948; and W-B, range (.932, .960), mean = .947, median = .949, resulted in coverage probabilities much closer to the true value of .950. The widths ranged from .421 to 2.366 with mean = 1.219 and median = 1.126. Six robust methods, LMS-B, LTS-B, S1S, MM-B, S, and ROB, led to over-coverage probabilities, range (.954, .997), mean = median = .977. Of the 36 conditions, 16 (or 44.4%), on average, fell within (.925, .975). The widths ranged from .399 to 1.449 with mean = .864 and median = .815. Three robust methods (lsfitci, LTS, and M) resulted in under-coverage probabilities, range (.394, .905), mean = .602, median = .489. Of the 36 conditions, 0 fell within (.925, .975). The widths ranged from .142 to 1.586 with mean = .576 and median = .299.

Summary of the Findings

Tables 44 and 45 present the mean, median, and range of the biases, RMSEs, Type 1 error rates, power rates, coverage probabilities and widths of the CIs for each of the six distributions of errors with the factors of the sample size and slope being aggregated. When all those criteria are considered, some important patterns of results are observed. First, OLS could produce, on average, a good point estimate of the slope value (as evidenced by the mean and median of bias and RMSE), but the range of those point estimates becomes much larger with MN, SH, AL, AH, and HE, as compared with other robust methods. OLS has a good protection of the Type 1 error for N, MN, SH, AL, and AH, but it leads to a very large Type 1 error for HE (mean = median = .378). The associated power rates are large for N (mean = .882, median = .949), but they drop substantially for MN, SH, AL, AH, and HE, as shown by the noticeably wider widths of the OLS CIs. Hence, even though OLS has both good protected Type I error rates and coverage probabilities of the true parameter value for N, MN, SH, AL, and AH, the precision of the point estimates is compromised and the widths of the CIs are too wide, leading to a noticeable decline in the observed power rates in potentially detecting any significant slope estimates, when they are indeed different from zero in the population. It is also noteworthy that OLS performed poorly for HE.

Second, HC3, HC4, and W-B appear to offer good adjustment when errors are HE. In particular, the means (or medians) of the Type I error rates were .056, .051, .055 (or .060, .054, and .053) for HC3, HC4, and W-B respectively. The coverage probabilities were also improved: the means (or medians) were .944, .949, and .947 (or .943, .948, and .949) for HC3, HC4, and W-B, respectively. On the other hand, HC3, HC4, and W-B still have the same limitation as in the conventional OLS-based CIs, meaning that the power rates were small, and the widths of the CIs were wide for MN, SH, AL, and AH. Hence, HC3, HC4, and W-B could potentially solve the issue of the Type I error and coverage of the true parameter for HE, but they may not be the most appropriate approach to use in practice when errors are non-normal (MN, SH, AL, AH, and HE).

Third, of the remaining robust methods, only M, GM, MM, MM-B, and ROB could be considered for MN, SH, AL, and AH, and only GM, E, and QUA could be considered for HE because of their superiority of the coverage probabilities that span the true parameter slope values. Comparatively, MM produced Type I error rates slightly larger than the criterion of .05, and it also consistently led to “no solution” for HE. Hence, it is not the most appropriate approach in practice. MM-B, a modified approach based on MM, seemed to overcome the no-solution issue with MM for HE, but its Type I error rates seem to fall in the conservative side of .05 (e.g., mean = .032, median = .031 for HE). ROB’s performance is also similar to MM-B’s performance, meaning that ROB behaved appropriately for N, MN, SH, AL, and AH, except for more conservative Type 1 error rates (mean = median = .028) for HE. Conversely, QUA seems to have good, protected Type 1 error rates for HE (mean = .047, median = .046), but they became noticeably smaller (or more conservative) for N, MN, SH, AL, and AH. E consistently produced smaller Type I error rates for all the six error conditions.

In sum, GM seems to have the best all-round performance. The Type 1 error rates remain slightly conservative for N, MN, SH, AL, and AH (means = .043, .042, .042, .040, and .041; medians = .043, .043, .041, .039, and .040, respectively) The power rates are reasonable and comparable to those obtained by M, MM, MM-B, and ROB for MN, SH, AL, and AH. When errors are HE, the Type 1 error rates are still appropriate (mean = .048, median = .047), and the power rates are desirable (mean = .625, median = .586) and comparable to those obtained by other robust methods.

Discussion

OLS regression is a widely employed statistical method in psychology (Anderson & Schumacker, 2003; Erceg-Hurn & Mirosevich, 2008; Haupt et al., 2014). However, its efficiency is often distorted in practice due to violated assumptions and the presence of outliers (Erceg-Hurn & Mirosevich, 2008; Micceri, 1989; Wilcox, 1998). When researchers have outliers in their data, robust regression may be a valuable alternative to handle outliers without causing other potential problems caused by other methods (e.g., changing the original construct or distribution). However, due to the existence of several types of robust regression estimators, which have their own strengths and weaknesses, it may be confusing and challenging for researchers to choose which robust regression method is appropriate for their research—without a clear guideline based on empirical evidence from a simulation study.

Implications of the Findings

The primary purpose of this study was to provide applied researchers with empirical evidence and guidelines to select more appropriate regression methods by comparing OLS and robust regression under different conditions. This simulation study suggests that when the normality assumption is met, OLS regression outperforms robust regression methods in terms of bias, RMSE, Type 1 error, power, coverage probabilities and confidence width of the CIs. This is because OLS regression achieves maximum efficiency when the normality assumption is met (Andersen, 2008; Field & Wilcox, 2017). Consistent with previous research (Andersen, 2008; Anderson & Schumacker, 2003; Mercer et al., 2015), the current results show that when the sample size per independent variable is large (n = 100), robust regression methods are quite comparable to OLS regression. That is, as the sample size increases, the efficiency of robust regression methods also increases, thereby paralleling the OLS regression method. Therefore, with sufficiently large samples, robust regression methods may be an appropriate alternative.

These research findings concur with those of other researchers (Andersen, 2008; Anderson & Schumacker, 2003; Brossart et al., 2011; Finger, 2010; Mercer et al., 2015; Sauvageau & Kumral, 2015; Yellowlees et al., 2016), and indicate that robust regression methods, in general, are better options than OLS regression when the normality and homoscedastic assumptions are violated. More specifically, when errors were non-normal (MN, SH, AL, AH, HE), HC3 and HC4 provide good adjustment for HE, but the associated power rates are still noticeably small and the widths of the CIs are wide, leading to a subpar approach in practice. Comparatively, GM is the most all-around and appropriate method in terms of the Type I error, power, coverage probability, width of the CI, as well as the precision of the point estimates. There are alternative robust methods which researchers could consider if they know that the errors are either with long tails (i.e., MM-B and ROB for MN, SH, AL, and AH) or heteroscedastic (QUA), if they would like to have slightly higher power with reasonably protected Type 1 error rates.

In addition to reporting the slope estimates, researchers often report and interpret the associated standard error and *p*-value to evaluate the sampling error or precision of the estimates. According to Wilcox and Keselman (2004), heteroscedasticity does not impact the slope but the standard error, which, in turn, influences the *p*-value. These findings provide empirical evidence that concurs with Wilcox and Keselman's explanation of heteroscedasticity. Therefore, although all regression methods perform well in terms of bias, only GM handles the Type 1 error rate effectively with desirable coverage probability of the CI. Therefore, GM seems to be an appropriate option for researchers to deal with heteroscedasticity.

Limitation and Future Directions

Although the results of this study provide a valuable guideline for future research regarding the use of appropriate robust regression methods, researchers need to consider the limitations of this study. First, although the Monte Carlo simulation tool allowed us to compare OLS and different types of robust regression methods under a variety of conditions, the findings were based on the simulated data. Therefore, like all other Monte Carlo simulation studies, the degree to which the findings may generalize to real data is uncertain and needs further examination. On the other hand, by including a variety of sample sizes, slopes, and error distribution conditions, based on previous studies, the research findings may serve as a guideline for future researchers when they select an appropriate robust regression method for their research. Future research may explore additional manipulated factors, such as larger sample size per predictor ($n > 100$) and other types of non-normal distributions; it could also compare OLS and robust regression methods based on real-world data in published studies. Second, this study shows that, broadly speaking, the GM estimation seems to perform appropriately across all the simulated conditions. Indeed, other studies (e.g., Wilcox, 2022) illustrated that it is overly simplistic for researchers to suggest one single estimator that is robust to all different levels or patterns of heteroscedasticity. Additional research is needed to further examine the details regarding the performance of various robust estimators under different levels and patterns of heteroscedasticity.

Conclusion

Due to the accessibility and advancement of computing power and the existence of free statistical software R and packages, researchers are readily able to conduct research using robust regression. When researchers suspect outliers in their data, but don't know the exact source, robust regression methods may be a valuable option to consider when addressing the specific outliers in their analysis. Based on the current research findings, researchers may be able to deal with outliers in their data efficiently using robust regression methods, especially if they use the GM-estimator. With sufficiently large sample sizes ($n = 100$), robust regression methods could be used by default instead of OLS regression without worrying about whether the error scores meet or violate the normality assumption. On the other hand, when the normality and homoscedasticity assumptions are met, OLS is found to offer small advantage in terms of slightly improved power, more precise width of the CIs, and protected Type I error as compared with other robust methods.

Funding: The authors have no funding to report.

Acknowledgments: This study is based on Jaejin Kim's honours research conducted at the University of Manitoba.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

For this article, the materials provided are the complete set of tables cited in this article (see Kim & Li, 2023a), and the code for the simulation study in R (see Kim & Li, 2023b).

Index of Supplementary Materials

Kim, J., & Li, J. C. (2023a). *Supplementary materials to "Which robust regression technique is appropriate under violated assumptions? A simulation study"* [Tables]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.13979

Kim, J., & Li, J. C. (2023b). *Supplementary materials to "Which robust regression technique is appropriate under violated assumptions? A simulation study"* [R code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.13980

References

Algina, J., Keselman, H. J., & Penfield, R. D. (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

Andersen, R. (2008). *Quantitative applications in the social sciences: Modern methods for robust regression.* SAGE Publications.

Anderson, C., & Schumacker, R. E. (2003). A comparison of five robust regression methods with ordinary least squares regression: Relative efficiency, bias, and test of the null hypothesis. *Understanding Statistics, 2*(2), 79–103. https://doi.org/10.1207/S15328031US0202_01

Brossart, D. F., Parker, R. I., & Castillo, L. G. (2011). Robust regression for single-case data analysis: How can it help? *Behavior Research Methods, 43*(3), 710–719. https://doi.org/10.3758/s13428-011-0079-7

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

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88*(423), 872–880. https://doi.org/10.1080/01621459.1993.10476352

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

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591–601. https://doi.org/10.1037/0003-066X.63.7.591

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour Research and Therapy, 98*(Supp. C), 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Finger, R. (2010). Revisiting the evaluation of robust regression techniques for crop yield data detrending. *American Journal of Agricultural Economics, 92*(1), 205–211. https://doi.org/10.1093/ajae/aap021

Finney, D. J. (2009). An interaction of medical and statistical ethics. *Journal of Medical Ethics, 35*(1), 51–52. https://doi.org/10.1136/jme.2008.025882

Fischer, A., Roodman, D., Zeilesis, A., Graham, N., Koell, S., Berge, L., & Krantz, S. (2023). *fwildclusterboot: Fast wild cluster bootstrap inference for linear models* (Version 0.13.0) [Computer Software]. R Project for Statistical Computing. https://cran.r-project.org/web/packages/fwildclusterboot/fwildclusterboot.pdf

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Prentice Hall.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology, 68*(1), 155–165. https://doi.org/10.1037/0022-006X.68.1.155

Handschin, E., Schweppe, F. C., Kohlas, J., & Fiechter, A. (1975). Bad data analysis for power system state estimation. *IEEE Transactions on Power Apparatus and Systems, 94*(2), 329–337. https://doi.org/10.1109/T-PAS.1975.31858

Haupt, H., Lösel, F., & Stemmler, M. (2014). Quantile regression analysis and other alternatives to ordinary least squares regression: A methodological comparison on corporal punishment. *Journal of Research Methods for the Behavioral, 10*(3), 81–91. https://doi.org/10.1027/1614-2241/a000077

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol.5., pp. 221–233). MR 0216620. Zbl 0212.21504.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics, 1*(5), 799–821. https://doi.org/10.1214/aos/1176342503

Huber, P. (1981) *Robust statistics.* Wiley. https://doi.org/10.1002/0471725250

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33–50. https://doi.org/10.2307/1913643

Krasker, W. S., & Welsch, R. E. (1982). Efficient bounded-influence regression estimation. *Journal of the American Statistical Association, 77*(379), 595–604. https://doi.org/10.1080/01621459.1982.10477855

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *American Statistician, 54*(3), 217–224. https://doi.org/10.2307/2685594

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods.* Wiley.

McCrone, P., Knapp, M., & Fombonne, E. (2005). The Maudsley long-term follow-up of child and adolescent depression. *European Child & Adolescent Psychiatry, 14*(7), 407–413. https://doi.org/10.1007/s00787-005-0491-6

Mercer, S. H., Lyons, A. F., Johnston, L. E., & Millhoff, C. L. (2015). Robust regression for slope estimation in curriculum-based measurement progress monitoring. *Assessment for Effective Intervention, 40*(3), 176–183. https://doi.org/10.1177/1534508414555705

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

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. *Measurement and Evaluation in Guidance, 6*(3), 157–163. https://doi.org/10.1080/00256307.1973.12022590

Mossakowski, K. N. (2011). Unfulfilled expectations and symptoms of depression among young adults. *Social Science & Medicine, 73*(5), 729–736. https://doi.org/10.1016/j.socscimed.2011.06.021

Olsson, D. M., & Nelson, L. S. (1975). The Nelder-Mead simplex procedure for function minimization. *Technometrics, 17*(1), 45–51. https://doi.org/10.1080/00401706.1975.10489269

Osborne, J. W. (2003). Notes on the use of data transformation. *Practical Assessment, Research & Evaluation, 8*(6), 1–7. https://doi.org/10.7275/4vng-5608

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should ALWAYS check for them). *Practical Assessment, 9*(6), 1–8. https://doi.org/10.7275/qf69-7k43

R Core Team. (2023). *R: A language and environment for statistical co*mputing. R Project for Statistical Computing. https://www.R-project.org/

Roodman, D., Nielsen, M. Ø., MacKinnon, J. G., & Webb, M. D. (2019). Fast and wild: Bootstrap inference in Stata using boottest. *Stata Journal, 19*(1), 4–60. https://doi.org/10.1177/1536867X19830877

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association, 79*(388), 871–880. https://doi.org/10.1080/01621459.1984.10477105

Rousseeuw, P. J., & Leroy, A. M. (2003). *Robust regression and outlier detection.* Wiley-Interscience.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdly, & D. Martin (Eds.), *Robust and nonlinear time series analysis* (pp. 256–272). Springer.

Sauvageau, M., & Kumral, M. (2015). Analysis of mining engineering data using robust estimators in the presence of outliers. *Natural Resources Research, 24*(3), 305–316. https://doi.org/10.1007/s11053-014-9254-8

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*(3), 699–714. https://doi.org/10.1177/001316447103100310

Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* McGraw-Hill.

Stromberg, A. J., Hössjer, O., & Hawkins, D. M. (2000). The least trimmed differences regression estimator and alternatives. *Journal of the American Statistical Association, 95*(451), 853–864. https://doi.org/10.1080/01621459.2000.10474277

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817–838. https://doi.org/10.2307/1912934

Wilcox, R. R. (1998). The goals and strategies of robust methods. *British Journal of Mathematical & Statistical Psychology, 51*(1), 1–39. https://doi.org/10.1111/j.2044-8317.1998.tb00659.x

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

Wilcox, R. R., & Keselman, H. J. (2004). Robust regression methods: Achieving small standard errors when there is heteroscedasticity. *Understanding Statistics, 3*(4), 349–364. https://doi.org/10.1207/s15328031us0304_8

Wilcox, R., & Xu, L. (2023). Regression: Identifying good and bad leverage points. *International Journal of Statistics and Probability, 12*(1), 1–8. https://doi.org/10.5539/ijsp.v12n1p1

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics, 14*(4), 1261–1295. http://www.jstor.org/stable/2241454

Yellowlees, A., Bursa, F., Fleetwood, K. J., Charlton, S., Hirst, K. J., Sun, R., & Fusco, P. C. (2016). The appropriateness of robust regression in addressing outliers in an anthrax vaccine potency test. *Bioscience, 66*(1), 63–72. https://doi.org/10.1093/biosci/biv159

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

EUROPEAN ASSOCIATION OF METHODOLOGY

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.