Statistics and Mathematical Research Journal
ISSN: 2997-6898 |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E32
Official Journal of Ethan Publication

A COMPARATIVE EVALUATION OF TELBS ROBUST REGRESSION

[1]Jane Marie Carter and [2]Michael Andrew Stevenson
[1]Department of Mathematical Sciences, Riverdale University, Riverdale, NY 10471, USA
[2]Mathematics Department, Coastal City University, Baytown, FL 32401, USA

Abstract:
Linear regression is a widely used statistical model for assessing the impact of explanatory variables on a response variable, with applications spanning across fields such as social sciences, environmental studies, and biomedical research. Conventionally, ordinary least squares (OLS) estimation has been the go-to method for regression analysis. However, the presence of outliers can significantly skew OLS parameter estimates. This paper delves into the intricacies of outliers, influential points, and leverage points in regression analysis, shedding light on their potential impact on parameter estimation. Outliers, whether in Y-space, X-space, or both, can distort the robustness of OLS estimations. Leverage points, often overlooked in standard residual plots, may exert substantial influence on regression coefficients. To address these challenges, a range of robust regression methods has been developed, including M estimation, MM estimation, LTS, S estimation, and the recently developed TELBS robust estimation. Section 2 provides an overview of these techniques. Robust regression approaches offer the means to estimate model parameters that either mitigate or exclude the influence of outliers, safeguarding against significant parameter distortions that OLS may encounter. Robust estimators are assessed in terms of breakdown point and asymptotic efficiency, crucial for evaluating the trade-off between robustness and data variability. While the application of robust regression has garnered traction in diverse fields, including sociology, political science, chemistry, and biology, its full adoption across all disciplines is an ongoing process.

Keywords: Linear regression, outliers, influential points, leverage points, robust regression.

1 Introduction

Linear regression is one of the most commonly used models for analyzing the effect of explanatory variables on a response variable. It has widespread application in various field of study, including social science, the environment, and biomedical research. The ordinary least squares (OLS) method has been generally used for regression analysis. However, OLS estimation of parameters is easily affected by the presence of outliers in the data. Outliers are observations that are far away from the main pattern of the data, while influential points and leverage points refer to the impact of removal of a point on the regression coefficients, in some sense the significance of exclusion of the given point. Outliers could be outlying in Y-space, X-space, or both. Usually, outliers outlying in X-space are also referred to as leverage points, such points do not always show up in the usual least square residual plots but may have significant influence on the regression coefficients. To remedy this problem, many robust regression methods have been developed that are not easily affected by the outliers including M estimation, MM estimation, LTS, S estimation, and a newly developed TELBS robust estimation. We give an overview of these methods in Section 2. In cases the data contains outliers, least squares method can be used to estimate the parameters after the outliers are identified and removed. Another way is using a robust regression approach other than OLS to estimate the model parameters without discarding any outliers. The robust regression methods provide a means to estimate model parameters that either diminishes or excludes the influence of outliers, which otherwise could have a significant impact on parameter values when using OLS method. In certain cases the outliers may represent a departure from the larger pattern that still has significance, while in other cases they may represent bad or invalid observations to be discarded. Two of the primary concerns we address in evaluating robust estimators are breakdown point and asymptotic efficiency, and these relate to a trade-off between robustness and variability in the data that is diminished when unusual observations are excluded or reweighted to be of less influence. These issues are discussed further below. Although application of robust regression would apply to analysis of data in a wide range of fields, from sociology and political science to chemistry and biology and beyond as described in Andrew, et al [2], while application of these methods has not yet fully caught on in all these fields.

Concepts of robustness have developed considerably in recent decades, such as addressed in Stigler [28], along with significant development in robust statistical tools and techniques, with the promise of ongoing development and new directions in development and application. Nevertheless we observe some studies where robust regression plays a central role, such as Kocak et al. [15], where linear regression is applied to estimate tissue resection weights in patients undergoing reduction mammaplasty. Mircean et al. [19] used Huber's M estimation to obtain an estimate of the ratio between expressions of specific proteins from two samples. We comment further below regarding examples of use of robust regression as an essential tool in bioinformatics, such as in the article of Xu et al. [34] analyzing the association between DNA copy number and gene expression through use of robust regression. An important aspect of robust regression includes identification of outliers and other leverage point. Distance measures, such as studentized residuals, Cook's distance, and the leverage statistic, h, also closely related to Mahalanobis distance, are often utilized for a quantitative assessment of which data points may be outliers or leverage points, Cook and Weisberg [6] and Rousseeuw and Hubert [25], and the need for such analytic methods to identify outliers or leverage points increases with large data sets and with high dimensional data, both of which are prevalent in many current applications. It is worthwhile to notice that in certain applications the outliers may be points of significant interest, such as addressed in Barghash, et al. [3], such as in data points in a data set of gene expression levels in cancer, which could represent a different subtype of the disease being studied. The distances in Barghash, et al. [3] are also applied to separate the extreme outliers interesting outliers that might be relevant for biological analysis. The article of Aguinis, et al. [1] also discusses the issues associated with robustness and identification of outliers, with attention to discussion associated to distinguishing error outliers, interesting outliers, and influential outliers, and how these may be handled differently.

This article builds on the article of Tabatabai, et al. [30] which introduced the new TELBS robust linear estimator, which displays both a high breakdown point and good asymptotic efficiency. This paper finds additional examples in biology and medicine in which robust methods, and particularly the TELBS method, can be applied to data sets from published works to improve the analysis over the non-robust OLS approach. Furthermore the computer simulations comparing TELBS to other methods made in Tabatabai, et al. [30] is extended to include comparison with Least Trimmed Squares (LTS) and S Estimation. Section 2 presents several important approaches to robust linear regression, culminating in the TELBS method. The following sections contain the extension of the work of [30] in comparing the performance of the TELBS method with these other methods. In Section 3, two real data sets from the medical field are analyzed using TELBS robust regression in comparison with other robust regression methods of S, LTS, M, and MM, and also compared to OLS and OLS with outliers removed. The computer simulation study to further investigate these methods in comparison with other robust methods is presented in Section 4. Finally, we give a summary and discussion in Section 5.

2. Robust Linear Regression Methods

We consider the standard multiple linear regression model given in the form of

$y=X\beta+\varepsilon$

where y is n by 1 response vector, $X=(x_{ij})$ is n by p design matrix of predictor variables, β is p by 1 vector of parameters, ε is n by 1 vector of random errors. The OLS estimate of parameter vector β is found by minimizing the sum of squared errors. For OLS the standard assumption is that residuals are independent and have identical normal distributions. Any departure from this strict assumption in the residuals can lead to difficulty in OLS estimation, such as addressed in Horn, et al. [12], MacKinnon [17], and Liu [16], and certain departures are commonly dealt with, such as heteroscedascity, presence of outliers, or combinations of multiple sources of error. Least squares necessarily penalizes outliers more based on the square of a large distance from the pattern, and the alternative of least absolute deviation (LAD) penalizes outliers less and provide some level of robustness to outliers. While OLS and LAD correspond to L^2 and L^1 norms for the residuals, other variations on relative weighting of residuals allows formation of other robust loss functions. Generalization to the concept of minimization of the robust loss function $\rho(x)$, often accomplished through use of the influence function $\psi=\rho'(x)$, provides the first robust method, M estimation, discussed further below.

In evaluating robust estimators, we primarily discuss two aspects of their performance, breakdown point and efficiency, described in Donoho and Huber [8]. Breakdown point relates to how much of the data can be corrupted before the robust estimator loses its effectiveness, i.e. the proportion of incorrect observations before the estimator gives incorrect estimates. Efficiency relates to the variance of the estimates with Absolute Relative Efficiency defined as Var_2/Var_1, and conceptually this measures how well the robust estimator compares with OLS for clean data, without outliers or leverage points.

In general, if larger residuals or outliers are reduced in significance or eliminated, this will diminish the variability and lower efficiency. Conceptually breakdown point be seen as having a limiting value of 0.5, as beyond this point it may not be possible to distinguish between underlying distribution and contamination, Rousseeuw and Leroy [23]. See the article Davies and Gather [7] for a more detailed discussion related to the limit of breakdown point, robust methods that achieve this limit, and associated assumptions. This article also includes development of associated ideas and concepts underlying the proofs of these results. We also mention the limitation of 0.5 for breakdown point is under the assumption that the robust estimator should have a unique solution. In relation to the applications to machine learning and computer vision, it is worthwhile to mention the alternative robust methods developed within this community to analyze data that has levels of contamination significantly higher than 50%. It is possible to reach level above 0.5 if do not require a unique solution. These methods are effective but computationally intensive and each has its own defect/weakness, such as discussed in Wang, et al. [32]. Breakdown point is one critical aspect of robust estimation, and various approaches have been developed to approach the 0.5 limiting value for the breakdown point. However, other important aspects of robustness should be considered, including efficiency, and Huber and Ronchetti [11] address issues of stability within robust estimation and the importance of efficiency, rather than solely focusing on breakdown point. We also mention that breakdown point is one of many considerations in applying robust methods, and it is of value to consider the comments of Cook, et al. [5] and Fox and Weisberg [9], relating to avoidance of very high

breakdown estimates, that very high breakdown estimates do not allow for diagnosis of model specification and require some level of certainty the data fit the underlying distribution.
Early robust estimators were higher efficiency and lower breakdown point, while new methods worked toward approaching 0.5 breakdown. After this stage,the goals shifted to moving toward increasing efficiency with a higher breakdown point. The earlier robust estimators to reach a high breakdown point had low efficiency. Selection or development of robust estimators seemed to offer trade-off between asymptotic efficiency and high breakdown point. For instance in the redescending influence functions discussed below, altering the tuning constant can increase rejection point and decrease efficiency. The work of Staudte and Sheather [29] addresses this trade-off and the importance of consideration of asymptotic efficiency together with breakdown point when performing robust estimation and selection of a balance appropriate to the context. In general the variation of the tuning constant varies the down-weighting of outliers, thus affecting the breakdown point and efficiency. Increasing the tuning constant tends to increase the breakdown point and decrease efficiency. However, modern methods have been developed that provide estimators with both a relative high breakdown point and high asymptotic efficiency, such as the MM and TELBS estimators described below. We also mention that sample size is an important consideration in relation to robust statistics, and the standard difficulties for linear regression with a small sample size, such as addressed in Chapelle, et al. [4], can be further compounded if there are outliers, contamination of data, or departure from distribution. and issues of small samples for robust estimation are well known, such as addressed in Imbens and Kolesar [13]. The problem of robust estimation with small sample size is also an important problem, and for these reasons the simulation section includes a relative small sample size, to demonstrate the value of the TELBS estimators in comparison to other estimators for both smaller and larger samples.

2.1 M Estimation

The concept of M estimation as weighted loss function with reweighted loss function ρ, as an extension of the concept of L_1 versus L_2 corresponding to different weight functions. This principle can be extended to more robust loss functions. Huber (1973) [10] introduced the M-estimate that minimize a function ρ of the errors. The objective function is given as

$$min\sum_{i=1}^{n}\rho(\frac{r_i}{\hat{\sigma}}) = min\sum_{i=1}^{n}\rho(\frac{y_i - x'_i\hat{\beta}}{\hat{\sigma}})$$

where σ is an estimate of scale. A reasonable ρ function should have several basic properties: non-negativity, zero penalty for residual of 0, symmetric, and monotonically increasing for residuals of increasing distance. These properties are given as the following:

$$\rho(r) \geq 0$$

, $\rho(0)$=0, $\rho(r)$=$\rho(-r)$, and $\rho(r_i)$≥$\rho(r'_i)$ for $r_i > r'_i$|. In addition we may often require the loss function to have a bounded influence function ψ, where ψ is the derivative of ρ. In general, influence function may be grouped in three categories: monotone, hard redescending, and soft redescending, with the Tukey bisquare function as an example of a hard redescender. The M estimate of the parameter can be obtained by taking partial derivatives with respect to β and setting them equal to 0. The system of normal equations are given by

$$\sum_{i=1}^{n}\psi(\frac{r_i}{\hat{\sigma}})x_i = 0$$

where ψ is the derivative of ρ, which will often be a bounded or redescending function. Note that the case of redescending influence function correspond to diminishing weights for increasing large residuals; this leads to the concept of rejection point based on shape of redescending function and the value of the tuning constant. This tuning constant affects where rejection point occurs, as well as affecting the balance between breakdown point and asymptotic efficiency for the estimator. Good examples to mention are the Haber and the Tukey bisquare influence functions, which are bounded and hard redescending, respectively. In these cases the tuning constant is the transition between levels of the influence and loss functions that determine where the influence function reaches its plateau or redescends to zero, respectively. Iteratively reweighted least square (IRLS) is one of the commonly used method to solve the nonlinear equations. In general, M estimate is fairly robust to the outliers in y-direction, however, it is not robust to leverage points (outliers in x-direction).

2.2 LTS Estimation

The Least Trimmed Squares (LTS) estimate was proposed by Rousseeuw (1984) [22] coupled with the Least Median of Squares (LMS) method, both of which were designed to achieve a high breakdown point. We note that the LTS method works by trimming a fixed percentage of the larger residuals. It thus has an inherent limitation of discarding a certain percentage of good data, or potentially missing a certain percentage of outliers, unless the number of outliers is known a priori. In effect, this usually leads to a reduced efficiency from reduced variance due to the discarded observations. Let $r_i = y_i - x'_i\beta$,i=1,…,n, the LTS estimate of parameter is given as

$$\hat{\beta} = \text{arg}min \sum_{i=1}^{h} r\,(2_i)$$

where $r_1^2 \leq r_2^2 \leq \cdots \leq r_n^2$ are the ordered squared residuals. Usually, h is defined in the range $n/2+1 \leq h \leq (3n+p+1)/4$, with n and p being sample size and number of parameters, respectively. We note that both the breakdown point and efficiency vary with h, as this adjusts the number of observations discarded. The trimming of the larger residuals makes LTS resistant to a significant number of outliers, and this method is analogous to a trimmed mean representing aspects of both the mean and median. LTS is considered as a high breakdown method with a breakdown point of 50%, i.e. to be resistent for a contamination of 50% of the data. The LTS has low asymptotic efficiency, as may be expected, since the loss of information in the trimming of a significant number of the data points will lead to a loss of variability and a lower efficiency.
We mention that LTS was one of the early methods to approach 0.5 breakdown point, its low efficiency is a limitation. While LTS has value as an early example of a robust estimator with high breakdown point, reaching practically 0.5 when h=n/2, its efficiency is low, approximately 7% when h=n/2 for the normal distribution. However, in practice LTS will sometimes be used as an initial estimate for other methods requiring a high breakdown point. The observation that the LTS estimator minimizes one form of a residual scale estimate leads directly to the S estimator, but consideration of minimization of a more efficient estimator of scale.

2.3 S Estimation

Because the adaptive weights of M estimates, associated toρ, are not invariant with respect to spread of the data, the concept of scale invariance leads to a new approach to estimation in S estimation. Here the underlying idea is to use M estimates in estimating the value of scale, S, and thus normalizing the role of scale in the relative down-weighting of outliers. This procedure of minimizing the scale estimate in the data also corresponds to minimization of the variance among the residuals. This S estimate was proposed by Rousseeuw and Yohai (1984) [26] and defined as

β =arg$min S(r_1(\beta),\ldots,r_n(\beta))$

where $r_i(\beta)$ is the i^{th} residual, the dispersion $S(\beta)$ is the solution of

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho\left(\frac{y_i - x'_i\hat{\beta}}{S}\right) = K$$

where $K = \rho(s)d\Phi(s)$ such that β and $S(\beta)$ are asymptotically consistent estimate of β and σ for the Gaussian regression model. This can be interpreted as expectation of loss function under a Gaussian distribution. Rousseeuw and Yohai [26] suggested a loss function as,

2

$\rho(x) = \frac{x^2}{2} - \frac{x^4}{2c^2} + \frac{x^6}{6c^4}, if|x| \leq c$, otherwise, $\rho(x) = \frac{c}{6}$.

The turning constant c controls the breakdown value and the efficiency of the S estimate. When c=1.548 and K=0.11995, the breakdown value of the S estimate is 50% and the asymptotic efficiency is about 29%. S estimation is usually considered as high breakdown and low efficiency method. The breakdown point and efficiency for S estimates vary with the tuning constant, and while this estimate can reach high breakdown point at relatively low efficiency, its breakdown point is fairly low at high efficiency, such as 10% breakdown point for 96.6% efficiency. This further illustrates the general principle of tradeoff between breakdown point and efficiency; however this is overcome in the next two estimators. Although the asymptotic efficiency for S estimates is somewhat improved over LTS, it is still considered as low efficiency. However, S estimates are often used in the initial step of the more refined robust estimators described below.

2.4 MM Estimation

MM estimation was introduced by Yohai (1987) [35]. It was the first estimate with a high breakdown (50%) and high efficiency under normal distribution assumption. MM estimation is based on a procedure by which an initial estimate with high breakdown point but low asymptotic efficiency can be restimated in an iterative procedure that increases efficiency while maintaining the initial breakdown point, Rousseeuw and Leroy [23]. This approach is based on holding fixed the scale from the initial estimate and applying another M estimate to the residuals. MM estimator includes three steps:

Step 1. Compute an initial consistent estimate $\hat{\beta}_0$ with a high breakdown point but possibly low efficiency (LTS estimate and S estimate are two kinds of estimates that can be used as the initial estimate). The commonly adopted loss function for S-estimate is given as

$$\rho(x) = \begin{cases} 3(\frac{x}{c})^2 - 3(\frac{x}{c})^4 + (\frac{x}{c})^6, & \text{if}|x| \leq c \\ 1 & otherwise \end{cases}$$

Step 2. Calculate the MM estimate of the parameters $\hat{\beta}$ that minimize the expression

$$\sum_{i=1}^{n}\rho\left(\frac{y_i - x'_i\hat{\beta}}{\hat{\sigma}_0}\right)$$

where σ_0 is the estimate of scale (standard deviation of the residuals) from first step.

Step 3. The final step computes the MM estimate of scale s which is the solution to the equation

$$\frac{1}{n-p}\sum_{i=1}^{n}\rho(\frac{y_i - x'_i\hat{\beta}}{s}) = 0.5$$

In Rousseeuw and Leroy [23] it is proven that the MM estimators inherit the high breakdown point from their initial estimate. However, since the breakdown point is based on the tuning constant in the initial two steps and the efficiency is based on a separate tuning constant in the final step, these can be made of be independent. Thus the MM estimate is the first case of high breakdown and high efficiency robust methods, and this is generally recognized as being very effective in dealing with multiple outliers and multiple leverage points. MM method yields an estimate with both high breakdown point and high asymptotic efficiency. This method is one of first successes in overcoming a trade-off between breakdown point and efficiency.

We mention briefly the tao robust method as being an alternative means of simultaneously achieving high breakdown point and high asymptotic efficiency, such as the tao robust methods in Yohai and Zamar [36] and Tabatabai and Argyros [31] and their applications in pattern recognition in Pern$i$ a-Espinoza, et al. [21] and Rusiecki [27]. This alternative method has the advantage of not requiring an initial estimate of scale.

2.5 TELBS Robust Linear Regression Method

The TELBS estimator is in a similar category to the MM estimation, with iterative steps from an initial estimate with high breakdown point, but this newer estimator also includes a variant approach to the influence function applying hyperbolic functions and other novel features. Tabatabai et al. [30] proposed this new robust linear regression method, TELBS method in 2012 and demonstrated its effective performance in comparison with other robust estimators, including MM. The TELBS estimate of parameter β is given by

$$\hat{\beta} = \mathrm{arg}min\sum_{i=1}^{n}\frac{\rho_\omega(t_i)}{L_i} \qquad (1)$$

where

$$\rho_\omega(x)=1-Sech(\omega x)$$

and ω, a positive real number, is called the turning constant, and its role in horizontal dilation/contraction of ρ_ω plays a role analogous to the tuning constants discussed above. The function $Sech(\cdot)$ is the hyperbolic secant function and t_i is defined by

$$t_i = \frac{(y_i - x'_i\hat{\beta})(1-h_{ii})}{\sigma} \qquad (2)$$

where σ is the error standard deviation, and h_{ii} is the diagonal element of the hat matrix of the form $H=X(X'X)^{-1}X'$,

where X is the design matrix. Define $M_j = Median\{|x_{1j}|,|x_{2j}|,\ldots,|x_{nj}|\}$ for $j=1,\ldots,p$. Define $L_i = \sum_{j=1}^{p} Max\{M_j, |x_{ij}|\}$. Usually, σ is unknown and it is suggested to use the estimator proposed by Rousseeuw and Croux [24], which is given by

$\sigma = 1.1926 Median(Median|r_i - r_j|), 1 \le i, j \le n$, (3)

where r_j is the j^{th} residual. Taking the partial derivatives of equation (1) with respect to the parameters and

setting them equal to zero results in the following system of equations:

$$\sum_{i=1}^{n}\frac{\psi_\omega(t_i)}{L_i}\frac{\partial t_i}{\partial \beta_i} = 0 \qquad (4)$$

where $\psi_\omega = \omega Sech(\omega x)Tanh(\omega x)$, which is the derivative of ρ_ω. We mention this influence function is a case of a soft redescending influence function with tuning constant ω. The tuning constant ω plays a role comparable to tuning constant c in the Tukey bisquare function given above. The weight w_i is defined as

$$w_i = \frac{\psi(t_i)(1-h_{ii})}{\sigma(y_i - x'_i\hat{\beta})L_i} \qquad (5)$$

Then the equation (4) can be written as

$$\sum_{i=1}^{n} w_i(y_i - x'_i\beta)x_i = 0$$

Denote the weight matrix by W, it is a diagonal matrix. The elements on the main diagonal are $w_1, w_2, \ldots, w_n$. Therefore, the estimate of the parameter β is given by

$$\beta = (X'WX)^{-1}X'W_y \qquad (6)$$

The following procedures are used to estimate the parameter.
Step 1. Set $\sigma^0 = 1$, calculate an initial estimate of vector β by minimizing the function given in (1).
Step 2. Calculate σ and weights w_i by using equation (3) and (5), then obtain the weight matrix W.
Step 3. Calculate β using equation (6).
Repeat steps 2 to 3 until convergence occurs.
For further details, see [30]. TELBS estimates of linear regression parameters have influence functions bounded in both the explanatory and the response variable direction. It has high breakdown point and high asymptotic efficiency. The tuning constant ω=0.405,0.525,0.628,0.721 correspond to 95%,90%,85%,80% efficiency, respectively.
In all examples and simulations considered in this study, TELBS method is evaluated under an asymptotic efficiency of
85%.
We note that TELBS is an important new robust estimator with high breakdown and high asymptotic efficiency, similar to the MM estimator. Comparisons with MM estimates and other robust estimates are made in the material in Tabatabai, et al. [30] and in the following sections containing applications of robust estimators to practical data from biology and computer simulations. These comparisons all suggest that TELBS performs on a level comparable to MM estimates, and with numerous examples where TELBS demonstrates a higher level of robustness.

3 Applications

To further study the performance of TELBS and compare it with other robust methods, OLS, S, LTS, M, MM, and TELBS were applied to two data sets. Tabatabai, et al. [30] proposed a new diagnostic measure for robust regression called S_h. We examine outliers by using three diagnostic measures: Cook's distance (CD), robust Studentized residual (SR), and S_h. A brief introduction of these measures is given below.
Cook's distance has been widely used for identifying outliers. Tabatabai et al. [30] suggested a robust Cook's distance using TELBS estimates of parameters, which is given by

$$CD_i = \frac{h_{ii}t_i^2}{p(1-h_{ii})^4}, \quad i = 1,2,\ldots,n$$

where p is the number of parameters, t_i is given by equation (2) and h_{ii} is the diagonal element of the hat matrix.
The Studentized residual using TELBS estimates of parameters has the form

$$SR_i = \frac{t_i}{\hat{\sigma}\sqrt{1-h_{ii}}}, \quad i = 1,2,\ldots,n$$

where σ is defined by equation (3).
In addition to considering the elements of the main diagonal of the hat matrix h_{ii}, Tabatabai, et al. [30] also recommended a new influential measure which is defined as

$$S_h(i) = \frac{h_{ii} - Median(h_{ii})}{\hat{\sigma}_h}, \quad i = 1,2,\ldots,n$$

where $\hat{\sigma}_h = 1.1926 Median(Median|h_{ii} - h_{jj}|), 1 \leq i,j \leq n$. Large value of $|S_h(i)|$ indicates the presence of an influential observation. This measure seems to be very good for identifying the leverage points based on the results in [30] and this study.
One example we studied is a brain and weight data that was taken from a larger data set in Weisberg [33] and Jerison [14]. It gives the brain weight and body weight of 65 animals. We used a logarithmic transformation (common log) for both variables, and a scatter diagram of the transformed data is given in Figure 1 (left). We can see there are 3 outlying points. Table 1 gives the values of the three measures for some of the observations using TELBS as a robust estimator of regression parameters. Observations 63, 64, 65 have large values for each measure and are identified as outliers. To investigate whether a larger brain is required to govern a heavier body, a linear regression model is used to fit the data with brain weight (y) and body weight (x). Each method was fitted to the data and the fitted lines for OLS and TELBS are given in Figure 1 (right). The OLS fit is pulled toward the outlying points and have the lowest slope.
We computed the parameter estimates, standard error and p-value using MASS package in R 2.12. However, the standard error and p-value are not available for S and LTS estimates. The result for each method is given in Table 2. Among five robust regression methods, MM and TELBS provide the closest estimates to OLS when the three outliers were removed. The small p-values indicate that body weight has a significant effect on brain weight, the higher the body weight, the larger the brain weight.
This example presents a clear linear relation in the log body weight versus log brain weight data, with three clear outliers that were also identified by each of the TELBS diagnostic measures: CD, SR, and Sh. Here the MM and TELBS methods are both very close to OLS with these outliers removed for both parameters, whereas each of S, LTS, and M had a larger distance for at least one of the parameters. This case is a good example where the MM and TELBS estimates provide a good representation of the linear relationship between log body weight and log brain weight, and the three removed outliers are a deviation from this pattern which significantly disrupt the OLS estimate, and to a lesser extend

Statistics and Mathematical Research Journal
ISSN: 2997-6898 |
Volume 11 Issue 1, January-March, 2023
Journal Homepage: https://ethanpub.online/Journals/index.php/E32
Official Journal of Ethan Publication

those of M, LTS, and S.

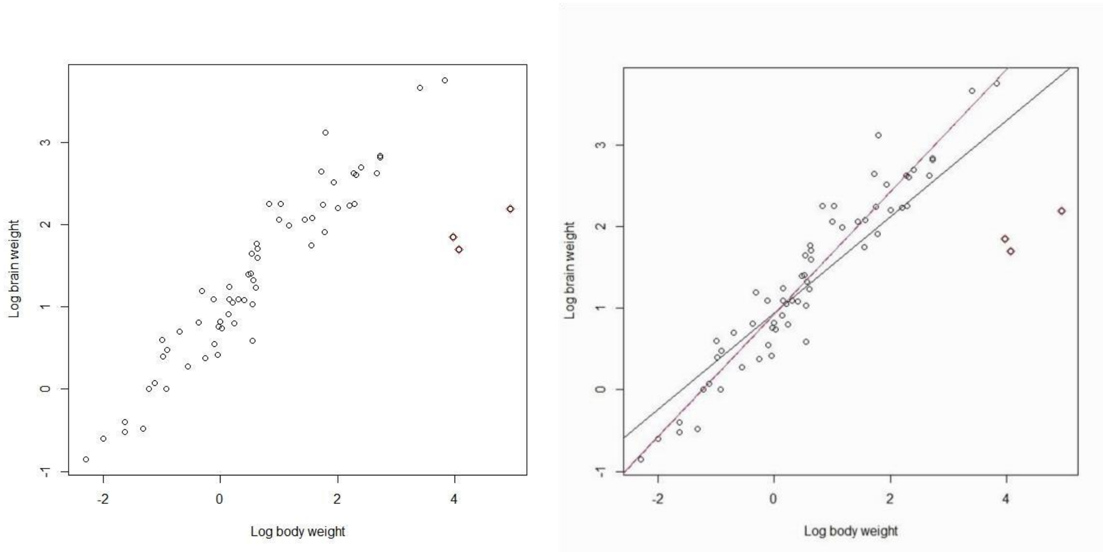

Figure 1: Left: Scatter diagram for brain weight data. Right: Scatter diagram with OLS (black line), TELBS fit (red line), and OLS fit with removal of 3 outliers (dashed line).

Table 1: Summary of diagnostic measures for brain weight data

Observation	CD	SR	S_h
1	0.0069	3.6013	-0.8091
2	0.0249	5.5704	0.1022
3	0.0011	-1.3180	-0.5288
4	0.0163	-3.3414	2.1828
⋮	⋮	⋮	⋮
62	0.0061	3.3907	-0.8374
63	2.2524	-24.5325	8.2490
64	1.7425	-22.2431	7.7345
65	4.2448	-25.6678	13.6624

Table 2: Summary of estimates for brain weight data for six comparison models

	Parameter	Estimate	Standard errors	P-value
OLS	Constant	0.9432	0.0704	<0.0001
	Log (Body weight)	0.5915	0.0412	<0.0001
OLS (removal of 3 outliers)	Constant	0.9271	0.0799	<0.0001
	Log (body weight)	0.7517	0.0464	<0.0001
S	Constant	0.8650	-	-
	Log (Body weight)	0.7470	-	-

LTS	Constant	0.8475	-	-
	Log (Body weight)	0.7713	-	-
M	Constant	0.9242	0.0462	<0.0001
	Log (Body weight)	0.6985	0.0270	<0.0001
MM	Constant	0.9196	0.0426	<0.0001
	Log (Body weight)	0.7460	0.0249	<0.0001
TELBS	Constant	0.9289	0.0435	<0.0001
	Log (Body weight)	0.7499	0.0255	<0.0001

Another example we considered is an interstitial lung disease (ILD) data which was used in Marubini et al. [18] and Narula et al. [20]. The data was collected to investigate the association between objective indicators of lung damage and severity of functional impairment in patients affected by ILD. Interstitial lung disease is a general category that includes many different lung conditions. All interstitial lung disease affect the interstitium, a part of the lung's anatomic structure. The response variable is forced vital capacity (FVC), that is the amount of air which can be forcibly exhaled from the lungs after taking the deepest breath possible. FVC is used to help determine both the presence and severity of lung diseases. The four factors studied in [20] that have impact on the FVC are age (in years), epithelial cells (EPIT: area fraction of epithelial cells/10000 pm^2 of alveolar tissue), cellular infiltration (CELL: total cellularity/10000 pm^2 of alveolar tissue), and Honeycombing (HONEY, score of honeycombing, zero to four).

In Marubini [18] and Narula [20], several outliers in the data were identified and two robust regression methods (C-M regression and the minimum sum of absolute errors regression) were applied to the data. We computed the values of the three diagnostic measures for each observation using TELBS estimators, the result for some observations is given in Table 3. The data set contains four outliers, observation 11 and 15 are outlying in the y-direction, observation 3 and 23 are leverage points that are outlying in the x-direction. 11 and 15 have the largest values in SR and relative large values in CD, while 3 and 23 have the largest values in S_h and CD.

Table 3: Summary of diagnostic measures for lung disease data

Observation	CD	SR	S_h
1	0.0004	0.0050	1.3729
2	0.0015	-0.0187	-0.6384
3	1.2315	-0.0505	7.6090
⋮	⋮	⋮	⋮
11	0.1700	-0.2369	-1.0388

⋮	⋮	⋮	⋮
15	0.3389	0.1757	0.7696
⋮	⋮	⋮	⋮
23	0.6848	-0.0484	6.8676
24	0.0010	0.0139	-0.3993

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

A multiple linear regression model is used to fit the data. For all the robust methods, the original data was used and the estimates of parameters, standard error and p-value for each method are given in Table 4. For OLS, results with and without outliers are also included in Table 4. R-square increased from 0.71 to 0.79 after the outliers were removed from the data which indicates a better fit to the data. The results indicate that the LTS estimate is clearly not suitable in this case as many of the parameter estimates are far from the values of both the estimates for OLS and OLS with outliers removed. In particular the parameter estimates for Constant and CELL are a different order of magnitude, far from those based on OLS. The parameters for HONEY and EPIT also display a significant difference compared to those for both OLS and OLS with outliers removed, and furthermore the estimates for the EPIT moves in the opposite (wrong) direction when applying LTS. Clearly in trimming of the data from LTS some important information in the data was lost along with the removal of the outliers. In fact, for this data set the S estimate displays some similar defects, though to a smaller extent. While the significant change in the parameter for Constant may be a concern, as it is no longer nearby the values for OLS and OLS with outliers removed, the significant change in the EPIT parameter in the opposite direction is a larger concern for this S estimate.

Table 4: Summary of estimates for lung disease data for six comparison models

	Parameter	Estimate	Standard errors	P-value
OLS	Constant	46.6539	11.2767	0.0006
	Age	0.6138	0.2364	0.0177
(R^2= 0.71)	EPIT	-0.0615	0.0174	0.0023
	CELL	107.7328	37.9187	0.0104
	HONEY	-10.6388	1.9359	<0.0001
OLS	Constant	52.7470	12.7672	0.0009
(removal of	Age	0.4540	0.1960	0.0351
4 outliers)	EPIT	-0.0590	0.0195	0.0086
(R^2= 0.79)	CELL	113.1916	53.9400	0.0532
	HONEY	-10.3710	1.5916	<0.0001
S	Constant	58.93483	-	-
	Age	0.40294	-	-
	EPIT	-0.06749	-	-
	CELL	108.56988	-	-
	HONEY	-10.38646	-	-
LTS	Constant	2.68466	-	-
	Age	0.52163	-	-
	EPIT	-0.09085	-	-
	CELL	334.44987	-	-
	HONEY	-8.50629	-	-
M	Constant	51.6893	11.2182	0.0002
	Age	0.4997	0.2351	0.0469
	EPIT	-0.0625	0.0173	0.0019

	CELL	112.4203	37.7221	0.0077
	HONEY	-10.6973	1.9259	<0.0001
MM	Constant	50.4173	10.7485	0.0002
	Age	0.5610	0.2253	0.0222
	EPIT	-0.0638	0.0166	0.0011
	CELL	108.2342	36.1424	0.0074
	HONEY	-10.8703	1.8452	<0.0001
TELBS	Constant	49.5598	8.2306	<0.0001
	Age	0.4599	0.1725	0.0077
	EPIT	-0.0431	0.0127	0.0007
	CELL	108.4482	27.67582	<0.0001
	HONEY	-10.1231	1.412981	<0.0001

Among the remaining estimates M, MM, and TELBS, all of these appear to be reasonable estimates, with results close to OLS and OLS with the outliers removed, while the TELBS estimates appear to be the best among these. Both M and MM estimates are comparable, and in this case the M estimates are in fact closer to the OLS with the outliers removed. Similarly to the S estimates, both M and MM had a change in estimates in parameters for both EPIT and HONEY that changed in the opposite (wrong) direction, a cause for concern. However these are a smaller order of magnitude than the others mentioned above. The TELBS estimate is relatively close in the estimates for all of the parameters, with the change of each of them in the same (correct) direction. Although some of the TELBS parameter estimates are closer to the estimates for OLS than to the estimates for OLS with outliers removed, all of the parameter estimates have the lowest standard error among the methods considered.

In this application to medical data of Marubini [18] and Narula [20] for interstitial lung disease, we see that the influence measures of CD, SR, and S_h associated to the TELBS robust estimator were useful in identifying outliers, as the stronger outliers in data points 3 and 23 displayed larger magnitudes of CD and S_h, while the outliers in data points 11 and 15 displayed larger magnitudes of SR. In application of robust estimators for the linear regression, TELBS performed favorably in comparison with the other methods considered, consistently with other examples considered in this article and in Tabatabai, et al. [30].

4 A Simulation Study

To further evaluate the performance of the TELBS estimates in comparison with M, MM, S, and LTS estimates, we conduct a simulation study under a small sample size (n=15), and a relative large sample size (n=30). We consider different contamination levels under various direction of contamination such as x-direction, y-direction, and both x- and y- directions. The simulation study is performed with R 2.12 and based on 5000 simulations. We consider a linear regression models with two covariates (x_1 and x_2) and generate both x_1 and x_2 and the random errors from a standard normal distribution with parameters 1, 3, and 3 for intercept and two covariates respectively. To evaluate the robustness of these estimates, we randomly chose 10%,20%,40% of the data and contaminated them by magnifying their size by a factor of 100, first in the direction of response variable (y), explanatory variables (both x_1 and x_2), then both the response and explanatory variables (y, x_1, and x_2). The results of these simulations are reported in Tables 5, 6, 7, 8, 9, and 10. In each case the estimators are evaluated in terms of Bias and MSE (Mean Square Error).

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

The bias was estimated by the equation $Bias = |\frac{\sum_{i=1}^{m}(\hat{\beta}_i)}{m} - \beta|$, where m is the number of simulations. The mean square error was estimated by $MSE = \frac{\sum_{i=1}^{m}(\hat{\beta}_i - \beta)^2}{m}$.
Table 5 and 6 give the results of Bias and MSE for each method for sample size of 15 and 30 respectively when the contamination is in the x- direction. For contamination in the y- direction, the results are given in Table 7 and 8, for sample size of 15 and 30 respectively. By examining the simulation results, we see that M estimation underperforms in all cases. It fails to give a close estimate of the parameters when the contamination level increases to 20% or higher. LTS, S, and MM estimation perform well in most cases except for a higher level of contamination, where they provide a relative large bias and MSE. In specific, S and MM fail to give a good estimate for the y-direction with a contamination level of 40% when the sample size is small. For x- direction, LTS, S and MM have relative large bias and MSE for both sample sizes when the contamination is 40%. TELBS outperforms all other methods in all six cases considered, it provides similar or smaller bias and MSE compared with other methods under each case. In general the performance of the estimators shows some level of difference for contamination in the x- direction and contamination in the y- direction; however this is not true for the TELBS estimator, which performs at a high level for all cases considered. In comparison with the other methods considered, this simulation has demonstrated the value of the TELBS robust estimator. Although there are a few cases where another estimator has numbers better than the TELBS estimator, TELBS has lowest value in majority of cases (rows in table), and in every case TELBS is nearby the lowest value. Furthermore TELBS performs most consistently throughout all cases considered, and there are many cases where it gives significantly better result than one or more of the others, especially in the cases of higher levels of contamination and contamination in the x-variable.
The best comparison with TELBS is MM, which is close in many cases and slightly better in some cases. MM slightly better for low and intermediate level contamination in the y- direction, in Tables 7 and 8, though the values are fairly close to TELBS. However, as the contamination increases to a higher level (40%), MM loses this slight advantage relative to TELBS, and in fact the values for MM are extremely poor in the case of the small sample size (n=15), as seen in Table 7. Also note that in some cases LTS performs well compared to TELBS, particularly in the case of high level of contamination in the y-direction for a small sample (n=15), where LTS had slightly lower values compared to TELBS, as seen in Table 7. However Table 8 shows that for a larger sample size, the MSE for estimation of these parameters is now smaller for TELBS. Furthermore TELBS performed significantly better than LTS in each of cases where the contamination is in the x- direction, and LTS had significantly worse scores than TELBS in cases of high levels of contamination in the x- direction, as seen in Tables 5 and 6. In cases of contamination in the x- and y- directions, in Tables 9 and 10, TELBS consistently show significantly lower MSE for the parameter estimates compared to LTS, while for high levels of contamination LTS was slightly better in regard to Bias of parameter estimates.
In summary, the simulation extends the trend seen in Tabatabai, et al. [30] and in the Applications section, which both emphasizing the value of the TELBS robust estimator in comparison with other robust estimation methods, such as M, S, LTS, and MM. This simulation revealed that TELBS yields best result in majority of cases and is consistently nearby the best result. Each of other estimators has at least one case with bad values, while TELBS is consistently the best estimate or near the best estimate for all the cases considered.

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

Table 5: Bias and MSE with contamination in the x-direction (n=15)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0063	0.0143	0.0433	0.0036	0.0133
	β_1	0.0308	0.0251	2.8605	0.0308	0.0293
	β_2	0.0453	0.0024	2.8787	0.0201	0.0041
MSE	β_0	0.3578	0.2009	1.4704	0.1013	0.1104
	β_1	0.5463	0.3102	8.3910	0.1811	0.1208
	β_2	0.4892	0.3025	8.4568	0.1765	0.1248
20% Bias	β_0	0.0013	0.0111	0.0615	0.0086	0.0004
	β_1	0.1122	0.0478	2.9603	0.0954	0.0243
	β_2	0.0333	0.0438	2.9595	0.0656	0.0004
MSE	β_0	0.3453	0.2407	1.4053	0.1277	0.1123
	β_1	0.6102	0.4066	8.7679	0.3716	0.1471
	β_2	0.5491	0.3729	8.7689	0.3598	0.1578
40% Bias	β_0	0.0184	0.0249	0.0116	0.0098	0.0074
	β_1	1.1986	2.0264	2.9694	0.5988	0.0945
	β_2	1.2098	2.0440	2.9698	0.4667	0.0846
MSE	β_0	0.7477	0.8525	0.8245	0.4445	0.1517
	β_1	3.8151	6.1377	8.8176	1.9756	0.2917
	β_2	3.7654	6.1337	8.8202	1.6455	0.3258

Table 6: Bias and MSE with contamination in the x-direction (n=30)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0006	0.0135	0.0106	0.0096	0.0003
	β_1	0.0175	0.0024	2.9442	0.0100	0.0048
	β_2	0.0193	0.0029	2.9463	0.0072	0.0068
MSE	β_0	0.1618	0.1091	0.6456	0.0430	0.0472
	β_1	0.1775	0.1213	8.6988	0.0646	0.0520
	β_2	0.1829	0.1289	8.7063	0.0622	0.0524
20% Bias	β_0	0.0001	0.0092	0.0449	0.0041	0.0097
	β_1	0.0058	0.0245	2.9683	0.0093	0.00003
	β_2	0.0007	0.0154	2.9684	0.0025	0.0001
MSE	β_0	0.1471	0.0939	0.5305	0.0508	0.0523
	β_1	0.1752	0.1216	8.8106	0.0744	0.0574
	β_2	0.1790	0.1297	8.8114	0.0729	0.0542
40% Bias	β_0	0.0146	0.0063	0.0305	0.0054	0.0036
	β_1	1.0612	1.7339	2.9695	1.7953	0.0105
	β_2	1.0295	1.7392	2.9697	1.7851	0.0114

MSE	β_0	0.1774	0.2492	0.3478	0.2372	0.0602
	β_1	3.2150	5.2153	8.8182	5.3454	0.0703
	β_2	3.2074	5.2251	8.8194	5.3484	0.0656

Table 7: Bias and MSE with contamination in the y-direction (n=15)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0143	0.0078	0.0204	0.0097	0.0002
	β_1	0.0415	0.0184	1.2148	0.0156	0.0034
	β_2	0.0247	0.0164	1.0158	0.0187	0.0149
MSE	β_0	0.2929	0.1992	22.7041	0.1042	0.1227
	β_1	0.3675	0.2245	144.0982	0.1153	0.1614
	β_2	0.4045	0.2210	145.7049	0.1239	0.1526
20% Bias	β_0	0.0143	0.0201	2.4941	0.0049	0.0137
	β_1	0.0118	0.0115	9.6629	0.0167	0.0037
	β_2	0.0146	0.0133	8.0754	0.0082	0.0043
MSE	β_0	0.2673	0.1701	401.0656	0.1025	0.1324
	β_1	0.3623	0.1972	1484.186	0.1278	0.1651
	β_2	0.3676	0.2014	1105.711	0.1332	0.1687
40% Bias	β_0	0.0097	1.4106	32.4270	4.2684	0.0013
	β_1	0.0018	4.4432	98.1562	18.6975	0.0071
	β_2	0.0203	4.6559	100.8857	19.6652	0.0214
MSE	β_0	0.1644	370.1686	4344.272	1001.424	0.1673
	β_1	0.2174	1181.942	16414.03	3050.492	0.2545
	β_2	0.2200	1293.672	17105.19	3449.291	0.2808

Table 8: Bias and MSE with contamination in the y-direction (n=30)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0183	0.0047	0.0355	0.0075	0.0156
	β_1	0.0088	0.0179	0.1092	0.0068	0.0061
	β_2	0.0207	0.0111	0.1175	0.0009	0.0024
MSE	β_0	0.1582	0.1002	0.0635	0.0435	0.0507
	β_1	0.1904	0.1140	0.0713	0.0465	0.0552
	β_2	0.1922	0.1103	0.0742	0.0431	0.0581
20% Bias	β_0	0.0077	0.0175	0.0038	0.0022	0.0077
	β_1	0.0023	0.0062	0.9126	0.0045	0.0055
	β_2	0.0026	0.0109	1.1691	0.0001	0.0076
MSE	β_0	0.1404	0.0896	24.1282	0.0475	0.0503

	β_1	0.1586	0.0979	47.0104	0.0491	0.0603
	β_2	0.1701	0.0921	108.3892	0.0498	0.0584
40% Bias	β_0	0.0096	0.0023	27.9896	0.0087	0.0061
	β_1	0.0047	0.0071	86.2142	0.0134	0.0074
	β_2	0.0010	0.0107	87.1350	0.0014	0.0054
MSE	β_0	0.0882	0.0811	2259.844	0.0727	0.0643
	β_1	0.0865	0.0823	11145.62	0.0834	0.0751
	β_2	0.1040	0.0912	11531.11	0.0809	0.0737

Table 9: Bias and MSE with contamination in both x and y-direction (n=15)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0122	0.0097	0.2276	0.0073	0.0065
	β_1	0.0088	0.0026	0.0878	0.0162	0.0078
	β_2	0.0091	0.0101	0.1919	0.0182	0.0161
MSE	β_0	0.3143	0.2312	4.3738	0.1174	0.1113
	β_1	0.4150	0.3584	21.6543	0.2469	0.1338
	β_2	0.4478	0.3159	22.2062	0.2289	0.1433
20% Bias	β_0	0.0257	0.0002	0.3366	0.0014	0.0106
	β_1	0.0003	0.0152	0.0514	0.0123	0.0212
	β_2	0.0248	0.0181	0.0562	0.0093	0.0283
MSE	β_0	0.3565	0.2052	2.0022	0.1325	0.1222
	β_1	0.4818	0.3436	5.2538	0.2807	0.1283
	β_2	0.5007	0.3403	6.5749	0.2593	0.1392
40% Bias	β_0	0.0167	0.0061	6.0666	0.0169	0.0265
	β_1	0.0532	0.0231	0.0448	0.0175	0.1012
	β_2	0.0228	0.0179	0.0371	0.0210	0.0276
MSE	β_0	0.3509	0.2635	160.056	0.2073	0.1496
	β_1	0.4819	0.3831	0.7726	0.3756	0.2257
	β_2	0.4859	0.3997	0.9163	0.3631	0.1975

Table 10: Bias and MSE with contamination in both x and y-direction (n=30)

	Parameter	LTS	S	M	MM	TELBS
10% Bias	β_0	0.0134	0.0047	0.1341	0.0006	0.0021
	β_1	0.0176	0.0065	0.0248	0.0116	0.0091
	β_2	0.0147	0.0154	0.0240	0.0092	0.0012
MSE	β_0	0.1674	0.1172	0.3124	0.0478	0.0474
	β_1	0.1827	0.1388	2.8842	0.0973	0.0581

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

	β_2	0.1759	0.1408	3.2812	0.0964	0.0565
20% Bias	β_0	0.0046	0.0107	0.3152	0.0011	0.0015
	β_1	0.0040	0.0165	0.0214	0.0030	0.0069
	β_2	0.0041	0.0054	0.0356	0.0029	0.0034
MSE	β_0	0.1579	0.1116	0.2796	0.0527	0.0491
	β_1	0.2466	0.1444	0.8944	0.1323	0.0548
	β_2	0.1935	0.1449	0.8763	0.1269	0.0563
40% Bias	β_0	0.0023	0.0005	6.2696	0.0075	0.0067
	β_1	0.0029	0.0078	0.0227	0.0012	0.0142
	β_2	0.0338	0.0073	0.0080	0.0221	0.0077
MSE	β_0	0.1286	0.0936	120.129	0.0802	0.0601
	β_1	0.1945	0.1595	0.3154	0.1551	0.0708
	β_2	0.1758	0.1606	0.3416	0.1545	0.0686

5 Discussion

This article presented the recently introduced TELBS robust linear estimator of Tabatabai, et al. [30] in comparison to other commonly used robust linear estimators, in particular M, LTS, S, and MM. The application of these linear robust estimators to two data sets, representing brain weight and interstitial lung disease (ILD) extends the results of Tabatabai, et al. [30] in illustrating the value of the TELBS estimator when applying linear regression to real data sets. In both cases the TELBS estimator performs favorably in comparison with these other robust linear estimators, with TELBS and MM at a comparable level of effectiveness. While both MM and TELBS yield results that are close to those of OLS with outliers removed, the TELBS estimator is slightly closer for both these cases. A more in depth comparison of the M, LTS, S, MM, and TELBS robust estimators through computer simulation yields an extended comparison of the performance of these estimators, and the results are comparable. Results of this simulation reveal deficiencies in M estimation, which fails to provide good estimates in some cases, especially when the sample size is small and the outliers are in x- direction. LTS, S, MM perform well in most cases except when the contamination level is high (40%). TELBS robust method performs well in all cases considered and outperforms other methods considered in this study as the percentage of outliers increases. It provides a flexible and powerful alternative to the practitioners in the field of robust linear regression.

References

Aguinis, H., Gottfredson, R. K., Joo, H. (2013). Best practice recommendations for defining, identifying, and handling outliers. Organization Research Methods. Sage 16:2, 270-301. Doi: 10.1177/1094428112470848.

Andrew, D. F., Bickel, P. J., Hampel, F. R. *et al.* (1972). Robust estimates of location: survey and advances. Princeton University Press, Princeton.

Barghash, A., Arslan, T., Helms, V. (2016). Robust detection of outlier samples and genes in expression datasets. J Proteomics Bioinform, 9, 38-48. Doi: 10.4172/jpb.1000387.

Chapelle, O., Vapnik, V. Bongio, Y. (2002). Model selection for small sample regression. Machine Learning 48, 9-23.

Cook, R. D., Hawkins, D. M., Weisberg, S. (1992). Comparison of model misspecifrication diasnostics using residuals from least mean of squares and least median of squares fits. J Amer Stat Assoc. 87:418, 419-424.

Cook, R. D., Weisberg, S. (1982). Residuals and Influence in Regression. New York, Chapman Hall.

Davies, P. L., Gather, U. (2007). The breakdown point examples and counterexamples. REVSTAT 5:1, 1-17.

Donoho, D. L., Huber, P. J. (1983). The notion of breakdown point. In A Festschrift for

Erich Lehmann (P.J. Bickel, K. Doksum, and J.L. Hodges, Jr., Eds.), Wadsworth, Belmont, CA, 157-184.

Fox, J., Weisberg, S. (2011). Robust regression in R. In: An R Companion to Applied Regression. (2nd ed.). Thousand Oaks, CA, Sage.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Stat, 1, 799-821.

Huber, P. J., Ronchetti, E. M. (2009). Robust statistics, second edition. New York: Wiley.

Horn, S. D., Horn, R. A., Duncan, D. B. (1975). Estimating heteroscedastic variances in linear models. J Amer Stat Assoc. 70:350, 380-385.

Imbens, G. W., Kolesar, M. (2012). Robust Standard Errors in Small Samples: Some Practical Advice. NBER Working Paper 18478.

Jersion, H. J. (1973). Evolution of the brain and intelligence. New York, Academic Press.

Kocak, E., Carruthers, K.H., McMahan, J.D. (2011). A reliable method for the preoperative estimation of tissue to be removed during reduction mammaplasty. Plast Reconsttr Surg, 127:3, 1059-1064.

Liu, R. Y. (1988). Bootstrap procedures under some non-iid models. The Annals of Statistics 16:4, 1696-1708.

MacKinnon, J. G. (2012). Thirty years of heteroscedasticity-robust inference. In Xiaohong Chen and Norman R.

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

Swanson, eds. Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis, pp. 437-461. Springer, New York.

Marubini, E., Orenti, A. (2010). Pinpointing outliers and influnetial cases in regression analysis: a robust method at work. Bioinformatics, 4:2, 95-105.

Mircean, C., Shmulevich, I., Cogdell, D. *et al.* (2005). Robust estimation of protein expression ratios with lysate microarray technology. Bioinformatics, 21:9, 1935-1942.

Narula, S. C., Saldiva, P.H.N. *et al.* (1999). The minimum sum of absolute errors regression: a robust alternative to the least squares regression. Statistics in Medicine, 18, 1401-1417.

Pern*i* a-Espinoza, A. V., Ordieres-Mer*e* , J.B., Martnez-de-Pis*o* n, F.J., Gonz*a* lez-Marcos, A. (2005). TAO-robust backpropagation learning algorithm. Neural Networks, 18:2, 191-204.

Rousseeuw, P. J. (1984). Least median of squares regression. J Amer Stat Assoc, 79, 871-880.

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

Rousseeuw, P. J., Croux, C. (1993). Alternative to the Median Absolute Deviation. J Amer Stat Assoc, 88, 1273-1283.

Rousseeuw, P. J., Hubert, M. (1999). Regression depth. J Amer Stat Assoc, 94:446, 388-402.

Rousseeuw P. J., Yohai, V. J. (1984). Robust Regression by Means of S Estimatorsin Robust and Nonlinear Time Series Analysis, edited by J. Franke, W. Hardle, and R.D. Martin, Lecture Notes in Statistics 26, Springer Verlag, New York, 256-274.

Rusiecki, A. (2013). Robust learning algorithm based on LTA estimator. Neurocomputing 120:23, 624-632. doi: 10.1016/j.neucom.2013.04.008.

Stigler, S. M. (2010). The changing history of robustness. The American Statistician, 64:4, 277-281.

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

Tabatabai, M., Eby, W., Li, H., *et al.* (2012). TELBS robust linear regression method. Open Access Medical Statistics, 2, 65-84.

Tabatabai, M. A., Argyros, I. K. (1993). Robust estimation and testing for general nonlinear regression models. Applied Mathematics and Computation, 58, 85-101.

Wang, Y. Rekaya, R. (2010). Lsoss: Detection of cancer outlier differential gene expression. Biomark Insights 5, 69-78.

Advanced Journal of Applied Mathematics and Statistics
Volume 10 Issue 1, January-March 2022
ISSN: 2837-4053
Impact Factor: 6.57
Journal Homepage: https://americaserial.com/Journals/index.php/AJAMS,
Email: contact@americaserial.com
Official Journal of America Serial Publication

Weisberg, S. (1985). Applied linear regression. (2nd ed.). New York: Wiley.

Xu, C., Liu, Y., Wang, P., *et al.* (2010). Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. Mol Cancer, 9:143.

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

Yohai, V.J., Zamar, R.H. (1988). High breakdown-point estimation of regression by means of the minimization of an efficient scale. J Amer Stat Assoc, 83:402, 406-413.