Doubly weighted m-estimator of linear regression model parameters

S. O. Ogundele[1,1] **and B. F. Ajibade**[2]

[1]Department of **Statistics**, Federal University of Petroleum Resources, Effurun, Delta State, Nigeria.
[2]Department of General Studies, Petroleum Training Institute, Effurun, Delta State, Nigeria.

Ordinary Least Squares (OLS) estimator produces the Best Linear Unbiased Estimate (BLUE) of the parameter of linear regression model if the assumptions of normality and constant variance of the error terms are satisfied. However, the assumption of the constant error terms across the entire observations is frequently violated by real life data. Due to the failure of OLS estimator for contaminated data, robust alternatives such as Least Absolute Deviation (LAD) method and M-estimators are proposed. M-estimators are robust to outliers in the y-direction but fail for x-outliers. To obtain M-estimator that is robust to outliers in both directions, weights were applied to two loss functions L_x and L_y to remove the effect of outliers in y and x directions. The method handles both the simple and multiple linear regression models and yields set of solutions that are unbiased and efficient. Comparative analysis of the performance of the proposed method with the existing methods indicates that the method competes favourably and are particularly more robust and efficient than other estimators considered when outliers lie on the X-direction and on both X and Y directions. The finite sample performance of the proposed method is studied using Monte Carlo simulation.

Keywords: M-estimator; Monte Carlo simulation; robust linear regression model

1. Introduction

Regression is a statistical methodology applied to relate a variable of interest, which is called the dependent variable or response variable, to one or more predictors (independent/regressors) variables. The objective of regression analysis is to build a regression model or prediction equation that helps us to describe, predict and control the dependent variable on the basis of the independent variables. Our interest, when predicting the dependent variable (Y) for a particular set of values of the independent variables (X_1, X_2, …, X_p), is to place a bound on the error of prediction so as to get predicted values with the smallest possible error.

Ordinary Least Squares (OLS) regression estimation relies on assumptions concerning the error terms of the model. OLS estimation assumes, among others, that the errors of prediction are independently and identically normally distributed, that is, [$r \sim N(0, \sigma^2)$]. Outlying observations, which is common in most life data often leads to violation of this assumption (Nevitt & Tam, 1998). It is well demonstrated that outliers in the sample data heavily influence estimates using OLS regression, sometimes even in the presence of one outlier (Rousseeuw & Leroy, 1987). Manimannan et al. (2020) studied the detection of outliers in linear and nonlinear regression models using standardized scores without the use of predicted values. The problem of contradiction between the statistical significance and real significance of regression parameters when using multiple linear regression analysis was studied (Yahya

[1] Corresponding Author; E-mail: ogundele.olaniyi@fupre.edu.ng

and Rezami, 2020). They presented an algorithm based on the simple and multiple coefficient of determination and the sum of averages to estimate multiple outliers when outliers are real. Raj and Kannan (2017) studied the detection and presence of outlying observations in simple linear regression model for medical data set.

If the assumption that the uncertainties (errors) in the data are uncorrelated and normally distributed are valid for the data at hand, then for most quantitative experiments, the method of least squares is the best linear unbiased estimator (BLUE) for extracting information from a set of data. The method is best in the sense that the parameters determined by the least squares analysis are normally distributed about the true parameters with the least possible standard deviations (Wolberg, 2006). The choice of estimation method under non-ideal conditions has been a long-standing problem for methodological researchers (Nevitt & Tam, 1998).

A regression procedure is said to be robust if the presence of contaminated data does not impede the procedure's ability to capture the general trend of the data. This idea is used to formulate a measure of the degree of robustness of estimator and it is called the breakdown point of a particular regression method. The higher the breakdown point the more resistant the estimator is to data contamination. The purpose of this study is to propose a Least Absolute Deviation (LAD) estimator that will be robust to outliers in both x and y directions while still retaining efficiency for Gaussian errors. Section 2 contains the materials and methods used in the study while sections 3 and 4 contain results and conclusion respectively.

2. Materials and Methods

`Assuming that the relationship between dependent variable (Y) and the independent variables (X) can be represented by the general linear regression model

$$y = X\beta + r \tag{1}$$

where β is a (p+1) by 1 column vector of parameters $(\beta = \beta_0, \beta_1, \beta_2, \cdots, \beta_p)$, $X = [x_{ij}]_{nX(p+1)}$ is the design matrix with n observations for p independent variables where the first column contains only element 1 and a column vector $y = (y_1, y_2, \cdots, y_n)$ of n dependent variable. The error terms associated with the model are represented by the column vector $r = (r_1, r_2, \cdots, r_n)$. We review briefly some of the existing methods used in estimating the parameters of the linear regression.

2.1 Ordinary least squares (OLS)

The LS criterion can be written as the value of $\hat{\beta}$ for which $L(\hat{\beta})$ is a minimum, where

$$L(\hat{\beta}) = \sum_{i=1}^{n} {r_i}^2(\hat{\beta}) \tag{2}$$

with r representing the OLS residuals. OLS estimator is optimal under the classical regression assumptions of independent and identically distributed normal errors, as it leads to parameter estimates that are BLUE.

2.2 Robust Regression Methods

The Least Absolute Deviation (LAD) methods for regression modelling minimises $L(\hat{\beta})$ where

$$L(\hat{\beta}) = \textstyle\sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{p} \hat{\beta}_j x_{ij} \right| \tag{3}$$

The problem is transformed into linear program problem which leads to linear programming algorithm for computation of LAD estimates. Portnoy and Koenker (1997) used interior point methods for solving linear programs with a new statistical pre-processing approach for L_1-type problems and proposed an interior-point linear programming implementation of the quantile regression estimator and its special case, the L_1 estimator. LAD is a low-breakdown regression estimator because its breakdown is less than 0.5.

Huber derived M-Estimator using a weight function ρ to down-weight any residuals resulting from outlying observations. The regression M-estimates is defined as solution $\hat{\beta}$ for which $L(\hat{\beta})$ is a minimum, where

$$L(\hat{\beta}) = \sum_{i=1}^{n} \rho\left(\frac{r_i(\hat{\beta})}{\hat{\sigma}}\right) \qquad (4)$$

where $\hat{\sigma}$ is some appropriately chosen estimate of σ. The choice is generally limited to robust measures of scale (Maronna et al., 2006). A robust estimate of scale that is frequently applied is the normalised median absolute deviation (MADN) given by

$$\hat{\sigma} = 1.4826\left[\underset{\forall i}{Med}\left(\left|r_i - \underset{\forall i}{med}\right|\right)\right], r_i \neq 0, i = ,2,\cdots,n \qquad (5)$$

The ρ -function down-weights observations with scaled residuals that are deemed too large in magnitude. Taking derivatives of (4) with respect to $\hat{\beta}$ leads to p normal equations to be solved to determine $\hat{\beta}$,

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\beta})}{\hat{\sigma}}\right) X_i = 0 \qquad (6)$$

where $\psi = \rho'$. These normal equations form a system of nonlinear equations and two methods frequently in use to solve the resulting nonlinear equations are Newton-Raphson and Iterated Reweighted Least Squares (IRWLS). IRWLS at convergence yields M regression parameter estimator that is equivalent to a Weighted Least Squares (WLS) estimator given by

$$\hat{\beta} = (X'WX)^{-1}X'Wy. \qquad (7)$$

W is an n x n diagonal matrix of observation weights $w_1, w_2, \cdots, w_n$ *where* $0 \leq w_i \leq 1, \forall i$. The weights are calculated using

$$w_i = \frac{\psi(r_i/\hat{\sigma})}{r_i/\hat{\sigma}} \qquad (8)$$

The weight function ψ determine the level of robustness of the estimator. Two common ψ functions in use are the Huber and bisquare ψ functions. The Huber ψ function is given as

$$\psi(r) = \begin{cases} -C_H, & if\ r < -C_H \\ r, & if\ |r| < C_H \\ C_{H,} & if\ r > C_H \end{cases} \qquad (9)$$

The parameter C_H is the tuning constant and $C_H = 1.345$ is proposed to achieve 95% efficiency under normal errors. The Huber weight function is given as

$$w(r) = \begin{cases} -C_H/r\,, & if\ r < -C_H \\ 1, & if\ |r| < C_H \\ C_H/r, & if\ r > C_H \end{cases} \qquad (10)$$

Another ψ function that is frequently in used is the bisquare ψ function. This is given by

$$\psi(r) = \begin{cases} 0, & if\ r < -C_B \\ r(1-(r/C_B)^2)^2, & if\ |r| < C_B \\ 0, & if\ r > C_B \end{cases} \qquad (11)$$

If the parameter C_B is the taking to be equal to 4.685 the estimator attain 95% efficiency under normal errors. The bisquare weight function is given as

$$w(r) = \begin{cases} 0, & if\ r < -C_B \\ (1-(r/C_B)^2)^2, & if\ |r| < C_B \\ 0, & if\ r > C_B \end{cases} \qquad (12)$$

The M-estimator is robust to outliers in y direction, but has no resistance against leverage points. Only one leverage point is enough to cause the M-estimator to breakdown and hence the breakdown of M-estimator is only $1/n$ (Rousseeuw & Leroy, 1987).

Least Trimmed Squares (LTS) estimator is based on the value of $\hat{\beta}$ which minimises $\sum_{i=1}^{h} r_{(i)}{}^2$, where $r_{(1)}{}^2 \leq \cdots \leq r_{(h)}{}^2$ are the squared residuals arranged in ascending order. The default value of h given by $h = [n/2] + 1$, yields a breakdown point of approximately 0.5 (Rousseeuw & Leroy, 1987).

Least Trimmed sum of absolute deviations (LTA) estimator is based on the value of $\hat{\beta}$ which minimises $\sum_{i=1}^{h} |r_{(i)}|$, where $r_{(1)} \leq \cdots \leq r_{(h)}$ are the absolute residuals written in ascending order. The default value of h given by $h = [(n + p + 1)/2]$, maximizes the breakdown of the resulting estimator (Hawkins and Olive, 1999).

MM-estimator derived by Yohai (1987), is a modified form of the Maximum Likelihood Type Estimator, M-Estimator, (Huber and Ronchetti, 2009), which estimates the regression parameters by determining the solution to the $(p + 1)$ equations

$$\sum_{i=1}^{n} \psi\left(\frac{r_i}{\hat{\sigma}}\right) x_{ij} = 0 \qquad (13)$$

$j = 0, \cdots, p$, where $\hat{\sigma}$ is a robust measure of variation based on the residuals using MADN. and

$$\psi(r_i, c) = \frac{r_i}{\hat{\sigma}}\left[\left(\frac{r_i}{c\hat{\sigma}}\right)^2 - 1\right]^2, if\ \left|{}^{r_i}/_{\hat{\sigma}}\right| \leq c; \qquad (14)$$

otherwise, $\psi(r_i, c) = 0$. The choice $c = 4.685$ leads to an MM-estimator with 95% efficiency compared to the least square estimator (Maronna et al., 2006).

2.3 Double weighted M-estimator (DWM)

In the criterion for M-estimator, the sum of the weighted vertical deviations is minimized and it is assumed that the sum of the horizontal deviations will not be significant enough to affect the estimation procedure. However, when there is a leverage point, the sum of the horizontal deviations is significant and could render the estimated parameter unacceptable as it is often biased and inefficient. We argue in favour of M-estimation procedure that weight down supposed outliers in both vertical and horizontal directions.

For simplicity, consider the simple linear regression model given by:

$$y = \beta_0 + \beta_1 x + r \qquad (15)$$

where y is the response or dependent variable, x is the independent variable or the predictor while r is the error term associated with the model. The purpose of regression analysis is to fit a model by finding the estimate $\hat{\beta}_0$ and $\hat{\beta}_1$ of the regression parameters β_0 and β_1 using sample data with minimum possible error.

The fitted model from the sample is given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{16}$$

Equation (9) can then be used to determine the fitted value $(\hat{y}_i)$ of y corresponding to x_i. The estimated error term $\hat{r}_i$ associated with the observation (x_i, y_i) is the difference between the observed response y_i and fitted response $\hat{y}_i$ given by:

$$\hat{r}_i = y_i - \hat{y}_i \tag{17}$$

The cut-off point (C) used to identify outlying observation is define as

$$C = k\hat{\sigma}, \tag{18}$$

where k is the turning parameter and $\hat{\sigma}$ is the robust estimate of the residual scale. The robust estimate commonly in use is the Median Absolute Deviation about the median (MAD) given by:

$$\text{MAD} = Med(|X_i - med(X)|),\ i = 1, 2, \cdots, n \tag{19}$$

To get a robust estimate that is consistent when the residuals follows the Gaussian distribution, the normalised median absolute deviation about the median (MADN) given by:

$$\text{MADN} = 1.4826\, Med(|X_i - med(X)|),\ i = 1, 2, \cdots, n, \tag{20}$$

is used.

If the residual is positive and greater than C for a particular outlying observation, the observation is adjusted to improve the fitness of the model as follows:

$$y_i^* = \hat{y}_i + C. \tag{21}$$

Observation with negative residual less than $-C$ are considered outlying and the fitness of the model is adjusted for improvement as follows:

$$y_i^* = \hat{y}_i - C \tag{22}$$

The iterative procedure are continued until the difference between successive estimates of the parameters is less than the tolerance value.

The procedure is repeated for x as the dependent variable and y as the independent variable using the model

$$x = \beta_0 + \beta_1 y + r \tag{23}$$

This double weighing will reduce the impact of supposed outliers and leverage points and the adjusted data are then used to build a model of y on x.

In multiple regression, after weighing of residuals using y as a function of all independent variables, multiple weighing is carried out by modelling each independent variable as a function of y and other independent variables. For model with two independent variables, say x_1 *and* x_2, we perform three weighing procedures using three different models:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + r \tag{24}$$

$$x_1 = \beta_0 + \beta_1 y + \beta_2 x_2 + r \tag{25}$$

$$x_2 = \beta_0 + \beta_1 y + \beta_1 x_1 + r \tag{26}$$

The adjusted data is then used to build the model of y on the other independent.

2.4 Breakdown point of regression estimator

Donohor and Huber (1983) introduced a finite sample breakdown point (FSBP) defined as follows:

Let $\Omega = \{w_i \epsilon R^p \ for \ i = 1, \cdots, n\}$ be a sample of size n. The breakdown point of an estimator T at Ω is given by

$$\varepsilon_n^*(T) = \frac{1}{n} max\left\{m: \overset{sup}{\widetilde{\Omega}_m} \left\|T\left(\widetilde{\Omega}_m\right)\right\| < \infty\right\}, \qquad (27)$$

where $\widetilde{\Omega}_m$ is any sample obtained from Ω by replacing any m of the points in Ω by arbitrary values. The largest *m/n* for which the property in (27) holds is the breakdown point of estimator T. The lowest breakdown point is 0 while the highest is 0.5 because at that point it is impossible to distinguish between contaminated and uncontaminated data.

2.5 Computational experiment

Monte Carlo Simulation is used to examine the robustness and efficiency of the proposed estimator. We performed experiment involving simple and multiple linear regression models. In the experiment, four levels of data contaminations (10%, 20%, 30% and 40%) were crossed with three sample sizes (20, 100 and 500) and this is repeated for both y and x variates. The assumed model for the simple regression is $y = 1 + 2x + r$ while for the multiple regression, the model is $y = 1 + 2x_1 + 5x_2 + r$. Hence the true parameters are 1, 2, and 5 for $\beta_0, \ \beta_1, and \ \beta_2$ respectively. Residuals for y outliers were generated using $N(10, 9)$ instead of the usual $N(0, 1)$ for non-outlying y variates. The x variates are generated as follows: $x \sim u(-1, 1)$ for simple regression, $x_1 \sim u(-1, 1), x_2 \sim u(-2, 2)$ for multiple regression. Outlying x variates were generated by using the uncontaminated x values to generate y and thereafter x and x_1 values are replaced by contaminated values generated from $x \sim u(10, 50)$ and $x_1 \sim u(10, 50)$ for simple and multiple regressions, respectively, using the idea of Rousseeuw & Leroy (1987).

All simulation programs were developed using R Statistical programming language (R Core Team, 2015). The function lm in the base package is used to obtain the estimate of regression parameters for the OLS estimator. The function rq in the package quantreg is used to obtain the estimate of LAD regression parameters using Portnoy and Koenker (1997) quantile regression algorithm. The function rlm in the package MASS is used to obtain the M estimate of regression parameters using Huber and Bisquare weighting functions. To obtain the estimate of regression parameters using MME algorithm (Yohai, 1987), we used the function lmrob in the package robust-base. The default value of the tuning parameter was used for all estimators.

Each simulation case was replicated $M = 1000$ times. The estimates of each estimator are calculated in each iteration and the mean of the M replicated estimates is given by

$$\hat{\beta}_j = \frac{\sum_{i=1}^{M} \hat{\beta}_{ji}}{M} \qquad for \ j = 0, 1, 2, \cdots, p \qquad (28)$$

is recorded for each estimator. Robustness of estimators is measured using absolute bias given as

$$AbsBias\left(\hat{\beta}_j\right) = \left|\beta_j - \hat{\beta}_j\right| \qquad for \ j = 0, 1, 2, \cdots, p \qquad (29)$$

A robust estimator has an estimate that is close to the actual parameter irrespective of the distortion in the distribution of the error terms. The lower the bias the more robust is the estimator. Efficiency of the estimators is measured using the MSE defined as

$$MSE\left(\hat{\beta}_j\right) = \frac{\sum_{i=1}^{M} (\beta_j - \hat{\beta}_{ji})^2}{M} \qquad for \ j = 0, 1, 2, \cdots, p \qquad (30)$$

and the variance of the estimator is defined as

$$Var(\hat{\beta}_j) = MSE(\hat{\beta}_j) - \left(Bias(\hat{\beta}_j)\right)^2 \qquad for\ j = 0, 1, 2, \cdots, p \tag{31}$$

The estimator with lowest MSE is the most efficient, the smaller the MSE the more efficient is the estimator.

3. Results

Due to limited space, the results presented represent the estimates of the estimators at the breakdown point. The summary of simulation results are presented in the Tables below.

Table 1: Bias for Simple and Multiple linear regression Estimates with no Outliers

Simple Regression, No Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	0.00317	0.00516	0.00367	0.00442	0.01608	0.00459
20	**Beta1**	0.00360	0.01166	0.00061	0.00238	0.15636	0.00229
100	**Beta0**	0.00271	0.00225	0.00233	0.00204	0.00166	0.00224
100	**Beta1**	0.00361	0.00648	0.00371	0.00396	0.11767	0.00372
500	**Beta0**	0.00047	0.00107	0.00024	0.00023	0.00159	0.00023
500	**Beta1**	0.00350	0.00337	0.00217	0.00239	0.18283	0.00243
Multiple Regression, No Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	0.00950	0.00243	0.01055	0.01205	0.03132	0.01005
20	**Beta1**	0.01506	0.02547	0.01571	0.01560	0.14069	0.01602
20	**Beta2**	0.00948	0.00432	0.01209	0.01269	0.05170	0.01213
100	**Beta0**	0.00168	0.00049	0.00124	0.00122	0.00566	0.00081
100	**Beta1**	0.00955	0.01052	0.00934	0.00904	0.11053	0.00764
100	**Beta2**	0.00061	0.00379	0.00153	0.00161	0.02774	0.00159
500	**Beta0**	0.00019	0.00141	0.00035	0.00038	0.00052	0.00038
500	**Beta1**	0.00015	0.00169	0.00070	0.00095	0.14026	0.00096
500	**Beta2**	0.00209	0.00163	0.00255	0.00255	0.03673	0.00256

We observed that when there are no outliers, all estimators are unbiased.

Table 2: Efficiency for Simple and Multiple linear regression Estimates with no Outliers

Simple Regression, No Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	0.05170	0.07437	0.05443	0.05614	0.05757	0.05553
20	**Beta1**	0.14687	0.22798	0.15625	0.16525	0.20383	0.16189
100	**Beta0**	0.00927	0.01421	0.00992	0.00994	0.01013	0.00993
100	**Beta1**	0.02476	0.03995	0.02656	0.02692	0.04633	0.02695
500	**Beta0**	0.00208	0.00345	0.00222	0.00223	0.00235	0.00223
500	**Beta1**	0.00602	0.00943	0.00627	0.00626	0.04134	0.00626
Multiple Regression, No Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	0.05860	0.08756	0.06052	0.06275	0.07719	0.06214

Table 2 Cont'd

20	**Beta1**	0.14930	0.23244	0.15352	0.16139	0.23033	0.15812
20	**Beta2**	0.03668	0.06159	0.03906	0.04145	0.05251	0.04036
100	**Beta0**	0.00924	0.01465	0.00965	0.00968	0.01255	0.00979
100	**Beta1**	0.02628	0.03903	0.02799	0.02813	0.05068	0.02863
100	**Beta2**	0.00838	0.01248	0.00879	0.00886	0.01167	0.00890
500	**Beta0**	0.00180	0.00305	0.00194	0.00196	0.00249	0.00196
500	**Beta1**	0.00554	0.00904	0.00593	0.00591	0.02774	0.00591
500	**Beta2**	0.00149	0.00228	0.00157	0.00158	0.00329	0.00158

OLS estimator is the most efficient and efficiency of all estimators increases with increase in sample size.

Table 3: Bias for Simple and Multiple linear regression Estimates with 30% Y-Outliers

Simple Regression, 30% Y-Outlier

Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	3.06678	0.40930	1.10097	0.14187	0.71431	0.13734
20	**Beta1**	0.43599	0.13121	0.27142	0.07431	0.16057	0.06321
100	**Beta0**	2.98413	0.39388	0.83380	0.07845	0.67795	0.09611
100	**Beta1**	0.36331	0.07543	0.14759	0.01271	0.18781	0.01633
500	**Beta0**	3.00362	0.39512	0.80030	0.07636	0.68272	0.09048
500	**Beta1**	0.03749	0.01305	0.02107	0.00158	0.05788	0.00199

Multiple Regression, 30% Y-Outlier

Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	2.72429	0.46143	1.21826	0.34948	0.60373	0.16153
20	**Beta1**	2.67519	0.73357	1.66849	0.55068	0.81808	0.22088
20	**Beta2**	0.14188	0.03680	0.03580	0.00345	0.25831	0.00325
100	**Beta0**	2.98557	0.39512	0.82745	0.07314	0.30508	0.09356
100	**Beta1**	0.11290	0.03556	0.05614	0.01425	0.14891	0.01546
100	**Beta2**	0.16071	0.03945	0.07191	0.01167	0.20081	0.01323
500	**Beta0**	2.98890	0.39664	0.79637	0.07704	0.33840	0.09146
500	**Beta1**	0.40418	0.07675	0.14312	0.01187	0.10745	0.01471
500	**Beta2**	0.05508	0.01189	0.02061	0.00393	0.14436	0.00405

Table 4: Efficiency for Simple and Multiple linear regression Estimates with 30% Y-Outliers

Simple Regression, 30% Y-Outlier

Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	10.71917	0.30926	1.86749	0.24872	0.70242	0.15918
20	**Beta1**	4.91288	0.52439	0.80624	0.48640	0.67968	0.49094
100	**Beta0**	9.18507	0.18180	0.77490	0.02551	0.49505	0.02913
100	**Beta1**	1.04221	0.10725	0.14493	0.07318	0.16122	0.07784
500	**Beta0**	9.06949	0.16121	0.65376	0.00978	0.47254	0.01227
500	**Beta1**	0.15061	0.01704	0.01904	0.01196	0.02205	0.01246

Multiple Regression, 30% Y-Outlier

Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	8.61498	0.41840	2.24990	0.80129	0.96599	0.18487

Table 4 Cont'd

20	Beta1	11.40943	1.42360	4.70397	2.04255	0.75716	0.58403
20	Beta2	1.18203	0.14450	0.29497	0.16311	0.40460	0.12547
100	Beta0	9.16629	0.18196	0.76732	0.02513	0.14590	0.02975
100	Beta1	0.72481	0.07770	0.10007	0.05293	0.13540	0.05681
100	Beta2	0.26087	0.02642	0.03686	0.01720	0.08384	0.01842
500	Beta0	8.98327	0.16281	0.64803	0.00987	0.12660	0.01247
500	Beta1	0.30838	0.02222	0.03785	0.01166	0.03242	0.01227
500	Beta2	0.04182	0.00452	0.00522	0.00300	0.02702	0.00312

The OLS estimator becomes biased and inefficient when the data contains 30% Y-outliers. Other estimators remain unbiased.

Table 5: Bias for Simple and Multiple linear regression Estimates with 20% X-Outliers

Simple Regression, 20% X-Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	0.05903	0.09191	0.05855	0.05833	0.25660	0.04733
20	Beta1	1.98274	1.97063	1.98260	1.98229	0.73554	0.77247
100	Beta0	0.04439	0.04779	0.04686	0.04660	0.27669	0.01310
100	Beta1	1.98650	1.98514	1.98595	1.98595	0.60007	0.22593
500	Beta0	0.04296	0.05020	0.04528	0.04523	0.27541	0.00045
500	Beta1	1.98623	1.98376	1.98543	1.98546	0.57465	0.00095
Multiple Regression, 20% X-Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	0.01846	0.04140	0.02341	0.02597	0.18724	0.00655
20	Beta1	1.98643	1.97920	1.98635	1.98625	0.05853	0.85997
20	Beta2	0.00068	0.00094	0.00006	0.00102	0.07465	0.01195
100	Beta0	0.05404	0.05410	0.05561	0.05558	0.24219	0.01372
100	Beta1	1.98608	1.98380	1.98529	1.98526	0.80357	0.28530
100	Beta2	0.00130	0.00056	0.00144	0.00145	0.05755	0.00032
500	Beta0	0.04199	0.04882	0.04454	0.04449	0.24084	0.00049
500	Beta1	1.98568	1.98270	1.98472	1.98475	0.74336	0.01225
500	Beta2	0.00155	0.00163	0.00107	0.00127	0.05617	0.00086

Table 6: Efficiency for Simple and Multiple linear regression Estimates with 20% X-Outliers.

Simple Regression, 20% X-Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	0.14115	0.24972	0.16120	0.16878	0.16035	0.12117
20	Beta1	3.93454	3.88884	3.93433	3.93329	0.91534	1.83499
100	Beta0	0.03139	0.05910	0.03466	0.03457	0.09042	0.01497
100	Beta1	3.94678	3.94189	3.94464	3.94466	0.41181	0.49897
500	Beta0	0.00732	0.01293	0.00814	0.00807	0.07844	0.00238
500	Beta1	3.94525	3.93554	3.94210	3.94219	0.34146	0.01215
Multiple Regression, 20% X-Outlier							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	0.15759	0.27431	0.18048	0.19136	0.20421	0.13292

Table 5 Cont'd

20	Beta1	3.94937	3.92281	3.94931	3.94915	0.54368	2.00664
20	Beta2	0.09826	0.16266	0.10917	0.11579	0.12147	0.08776
100	Beta0	0.03102	0.05637	0.03471	0.03458	0.07761	0.01733
100	Beta1	3.94515	3.93664	3.94207	3.94196	0.71207	0.60698
100	Beta2	0.01696	0.03146	0.01886	0.01877	0.01971	0.01127
500	Beta0	0.00715	0.01356	0.00806	0.00800	0.06158	0.00255
500	Beta1	3.94305	3.93135	3.93924	3.93936	0.56434	0.01564
500	Beta2	0.00359	0.00657	0.00395	0.00392	0.00630	0.00191

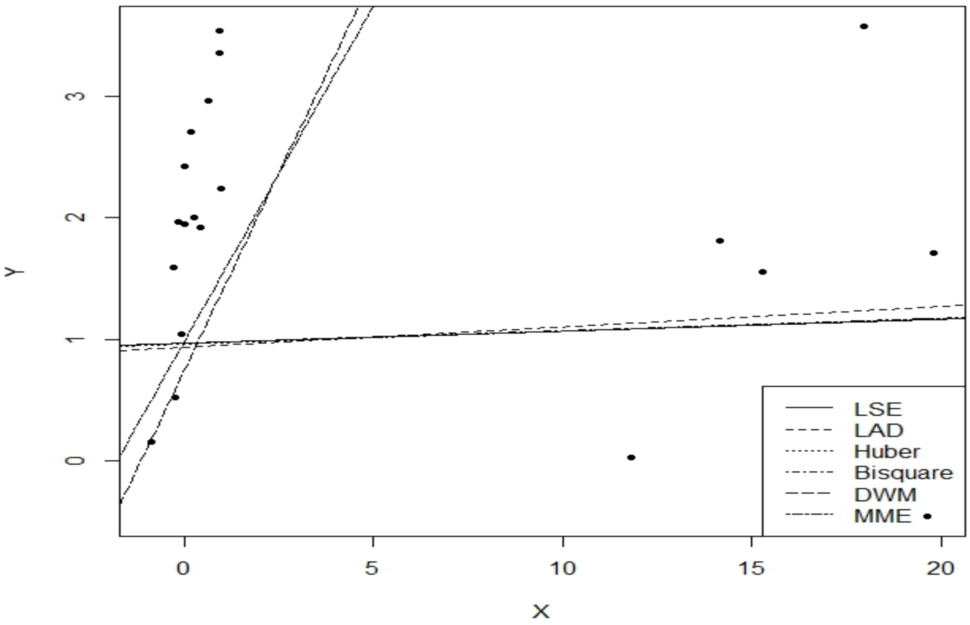

Figure 1: Scatter Plot and Fitted Line for Simple and Multiple linear regression Estimates with 20% X-Outliers

DWM and MME remain unbiased when the data contain 20% X-outlier other estimators are bias and inefficient. However, DWM is the only efficient estimator when the sample size is 20 (small sample) for both simple and multiple regressions.

Table 7: Bias for Simple and Multiple linear regression Estimates with 10% Y-Outlier and 10% X-Outlier

Simple Regression, 10% Each XY-Outliers							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	1.02787	0.08499	0.18938	0.00488	0.08193	0.01773
20	Beta1	2.04590	1.95960	1.99163	1.97848	0.74978	0.26895
100	Beta0	1.04492	0.11777	0.18477	0.02100	0.06552	0.01131
100	Beta1	2.04495	1.98163	1.98923	1.97585	0.80765	0.00691
500	Beta0	1.06856	0.14194	0.20181	0.00323	0.07682	0.01737
500	Beta1	2.04436	1.98146	1.98852	1.97537	0.81276	0.00134
Multiple Regression, 10% Each XY-Outliers							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	Beta0	1.10292	0.14618	0.21761	0.01214	0.13584	0.01064

Table 7 Cont'd

20	**Beta1**	2.06237	1.80525	1.92535	1.82684	0.91541	0.26519
20	**Beta2**	0.76651	0.14568	0.21926	0.03888	0.74935	0.00098
100	**Beta0**	1.04873	0.09555	0.16547	0.02074	0.28864	0.00517
100	**Beta1**	2.03793	1.86664	1.91641	1.79470	0.68196	0.03813
100	**Beta2**	0.95796	0.16635	0.23889	0.02995	0.89775	0.00022
500	**Beta0**	1.05692	0.09699	0.16619	0.01679	0.33414	0.00827
500	**Beta1**	2.04747	1.87476	1.93114	1.70920	0.57859	0.04547
500	**Beta2**	0.98999	0.17778	0.25051	0.03036	0.91178	0.00058

Table 8: Efficiency for Simple and Multiple linear regression Estimates with 10% Y-Outlier and 10% X-Outlier

Simple Regression, 10% Each XY-Outliers							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	1.67879	0.28766	0.20959	0.16583	0.12154	0.09361
20	**Beta1**	4.19404	3.84915	3.97240	3.92036	0.01520	0.85410
100	**Beta0**	1.21631	0.06923	0.06908	0.03215	0.02538	0.01470
100	**Beta1**	4.18353	3.92892	3.95822	3.90514	0.72689	0.06306
500	**Beta0**	1.16591	0.03099	0.04775	0.00667	0.01011	0.00310
500	**Beta1**	4.17972	3.92660	3.95447	3.90234	0.67541	0.00796
Multiple Regression, 10% Each XY-Outliers							
Sample Size	**Beta**	**OLS**	**LAD**	**Huber**	**Bisquare**	**DWM**	**MME**
20	**Beta0**	2.34868	0.31183	0.29099	0.18614	0.63277	0.09680
20	**Beta1**	4.41254	3.47838	3.89309	3.69184	0.68659	0.79067
20	**Beta2**	1.88901	0.27888	0.29146	0.15608	0.28216	0.07860
100	**Beta0**	1.30931	0.06620	0.06626	0.03158	0.21356	0.01402
100	**Beta1**	4.18336	3.53647	3.71726	3.33020	0.29309	0.07904
100	**Beta2**	1.16530	0.07129	0.09537	0.02793	0.95906	0.01114
500	**Beta0**	1.15857	0.02004	0.03586	0.00635	0.13506	0.00309
500	**Beta1**	4.19727	3.52358	3.73722	2.97265	0.49465	0.02459
500	**Beta2**	1.02439	0.04017	0.06993	0.00596	0.85988	0.00213

DWM and MME remain unbiased and efficient when the data contain 10% Y-outlier and 10% X-outlier other estimators are bias and inefficient.

4. CONCLUSION

Generally, when there are no outliers in the data, all estimators are unbiased and OLS is the most efficient estimator. The Monte Carlo simulation results agree with the classical results that OLS estimator is the Best Linear Unbiased Estimator (BLUE) under this condition. DWM and MME are unbiased and efficient while OLS, LAD, Huber and Bisquare estimators are bias and inefficient when there are leverage points and when contamination exist in both X and Y directions for large samples. However, DWM is the only efficient estimator for small sample when the data contains 20% X-outliers.

References

Afrah, Y. and Rezami, A.L. (2020). Effect of outliers on the coefficient of determination in multiple regression analysis with the application on the GPA for student, *International Journal of Advanced and Applied Sciences*, 1 7(10), 30 – 37.

Donoho, D.L. and Huber, P.J. (1983). The notion of breakdown point, In: *A Festschrift for Eric Lehmann*, Bickel P. J., Doksum K. A. and Hodges J. L. (Eds.), 157 – 184, Wadsworth, Belmont. CA.

Hawkins, D.M. and Olive, D. (1999). Applications and algorithms for least trimmed sum of absolute deviations regression, *Computation Statistics & Data Analysis*, 32, 119 – 134.

Huber, P.J. and Ronchetti, E.M. (2009). *Robust Statistics*, Second Edition, John Wiley & Sons Inc., New York.

Manimannan, G.M., Salomi, Priya R.L. and Saranraj, R (2020). Detecting outliers using R Package in fitting data with linear and nonlinear regression models, *International Journal of Scientific and Innovative Mathematical Research*, 8(4), 1 – 13.

Maronna, R.A., Martins, R.D. and Yohai, V.J. (2006). *Robust Statistics: Theory and Methods*, John Wiley & Sons Ltd, West Sussex.

Nevitt, J. and Tam, H.P. (1998). A comparison of robust and nonparametric estimators under the simple linear regression model, *Multiple Linear Regression Viewpoints*, 25, 54 – 69.

Portnoy, S. and Koenker, R. (1997). The Gaussian Hare and the Laplacian Tortoise: compatibility of squared-error versus absolute-error estimators, *Statistical Science*, 12(4), 279 – 300.

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

Stephen, R.S. and Senthamarai, K.K. (2017). Detection of outliers in regression model for medical data, *International Journal of Medical Research & Health Sciences*, 6(7), 50 – 56.

Wolberg J. (2006). *Data analysis using the method of least squares*, Springer-Verlag, Berlin Heidelberg.

Yohai, V.J. (1987). High breakdown-point and high efficiency estimates for regression, *Annals of Statistics*, 15(2), 642-656.