Abdulhamid Audu, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Comparative Study of Some Methods of Estimating Parameters of Linear Regression in Presence of Multicollinearity and Outlier

Warha, Abdulhamid Audu[1], Akeyede, Imam[2], Grema Modu Bako[3]
[1]Department of Statistics, Ramat Polytechnic Maiduguri, Borno State Nigeria
[2]Department of Mathematics Federal University Lafia, PMB 146 Lafia, Nigeria
[3]Department of Statistics, Ramat Polytechnic Maiduguri, Nigeria

Abstract- Classical least squares method for estimating regression models consisting of minimizing the sum of the squared residuals. Some of the assumptions of Ordinary least squares method (OLS) is that there is no correlations (multicollinearity) and extreme values (outliers) between the independent variables. Violation of these assumptions arises most often in regression analysis and can lead to inefficiency of the least square method. This paper therefore determined the efficient estimator between Least Absolute Deviation (LAD) and Weighted Least Square (WLS) in multiple linear regression models at different levels of multicollinearity and outlier in the explanatory variables. Simulation technique were conducted using R Statistical software, to investigate the performance of the two estimators under violation of assumptions of lack of multicollinearity and outliers. Their performances were compared at different sample sizes. Finite properties of estimators' criteria namely, mean absolute error, absolute bias and mean squared error were used for comparing the methods. The best estimator was selected based on minimum value of these criteria at a specified level of multicollinearity, outlier and sample size. The results showed that, LAD was the best at different levels of multicollinearity and outlier and was recommended as alternative to OLS under this condition. The performances of the two estimators decreased when the levels of multicollinearity and outliers was increased.

Keywords: LAD, WLS, Multicollinearity, Outlier.

I. INTRODUCTION

Linear Regression analysis is used to study the relationship between a single variable Y, called the response variable, and one or more explanatory variable(s), X1, X2, ..., Xp, by a linear model. One of the assumptions of Linear Regression model is that of independence between the explanatory variables (i.e. no multicollinearity). Violation of this assumption arises most often in regression analysis. Among methods used in detecting the presence of Multicollinearity is variance inflation factor (VIF) (Ajiboye et al, 2016). In the situation where the assumptions of the linear regression are not valid, many estimation methods have been proposed; Stein Estimator by Stein (1956), Liu Estimator by Liu (1993) and Ridge Estimator proposed by Hoerl and Kennard (1970) which is more efficient than OLS when there is collinearity in two or more explanatory variables.

The study therefore, determines efficient estimators among Least Absolute Deviation (LAD) and Weighted Least Square (WLS) Estimators in multiple linear regression models when there is correlation and extreme value(s) in the explanatory variables which are referred to as multicollinearity and outlier respectively. Their performance was compared for different sample sizes. Naturally, one would prefer best estimators which are fully efficient. Preferably, these estimators should also be robust to plausible deviations from an assumed model.

A common problem in regression models is outliers, which produces undesirable effects on the least squares estimators. Many regression estimation techniques have been suggested to deal with this problem. The majority of such techniques are developed from the classical least squares. Some other robust approaches have been investigated in the regression both on theoretical and empirical grounds. However, the best estimator(s) of regression models that contain some proportion of outliers and levels of multicollinearity have not received much attention in the context of time series and econometrics. The study therefore examined the effect of Outliers and Multicollinearity on performance of Least Absolute Deviation (LAD) and Weighted Least Square (WLS) Estimators in multiple linear regressions.

Problem of Multicollinearity

Some researchers are faced with a number of problems that arise because of the non-experimental nature of the discipline. Some of these problems arise because the researcher has to observe both the dependent and independent variables. This is in contrast to the position of the pure scientist who, in the experimental setting, can set the values of each of the explanatory variables and then observe their resultant effects on the dependent variable (Adnan et al, 2006). As long as he takes enough care at the planning stage, he would be able to estimate the effect of each independent variable precisely. In the social sciences situations, many of the explanatory variables show little variations while others show variations that are systematically related to variations in the other explanatory variables. This is the problem of multicollinearity. The basic regression method makes an explicit assumption that two or more of the explanatory variables do not have a perfect or almost perfect linear relationship. When this assumption breaks down then there is the problem of Multicollinearity (Akeyede et al, 2017).

Multicollinearity can therefore be defined as a measure of the degree of the linear relationship between two or more of the explanatory variables in a regression model. Thus the question to ask is the degree of the relationship rather than the existence of that linear relationship. The problem of multicollinearity with economic series is not whether it is present or not but finding out its severity. Many techniques have been proposed for doing this ranging from the traditional ones to the more scientific ones. Frisch (1934), as noted by Valentine (1969) was one of the first researchers to face the problem of detecting collinearity in a set of data. He proposed "a bunch map analysis" to do this. This is rarely practicable because of the computational burden. However regressions on subsets of the explanatory variables as well as on the full set may give useful information.

Base on simulation study, performance of Least Absolute Deviation and Weighted Least Square estimators have been evaluated by many researchers (Muniz and Kibria, 2009; Khalaf and Shukur, 2005; Alkhamisi et al, 2006; Alkhamisi and Shukur, 2008) on multiple linear regression especially when there is Multicollinearity. Additionally, they considered a number of regressors and used the MSE as the performance measure. In most of the studies when regression analysis was employed, observations were presumed to be equally and independently distributed (iid), despite that the iid assumption is much powerful in real-life contexts.

Problem of Outlier

The outliers are values which seem either too large or too small as compared to rest of the observations (Gumbel, 1960). An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969).

Barnett and Lewis (1984) give the following definition of an outlier. An outlier is a set of data in an observation that appears to be inconsistent with the remainder of that set of data. This definition identifies two aspects of outlier problem: the prior decision to group data together the apparent inconsistency that result. One possible solution to the outlier problem will be to rethink how we have group the data. May be the outlier is providing clues to the existence of another previously unrecognized sub population. Another possible solution will be to revisit why it appears inconsistent may be we have faulty underlying assumptions about how the data should behave (Barnett and Lewis, 1984).

The LAD reduces the effects of outliers among the Y values but it does not protect against outliers among the X values (leverage points), which have a large effect on the model fit. Hence, it is a robust with respect to outliers in the y-direction and it is not robust with respect to the x-outliers (Wilcox, 2010). The breakdown point of the least absolute value

(LAV) is 1/n. Thus the effect of one x-axis data outlier will change the regression line, causing the line to pass through the outlying data point.

The WLS regression is sensitive to the presence of outliers. If potential outliers are not properly addressed, they will definitely affect the parameter estimation and other aspects of a weighted least squares analysis (Habshah et al, 2014). Outliers may have a significant impact on the results of standard methodology for time series analysis, therefore it is important to determine a suitable method for estimating data that contain such outlier(s). However, we would like to determine the best methods for estimating regression model that contain outlier(s).

Other problems in regression analysis include the problem of outlier and leverage points. An outlier is an observation that is distant from other observations. Leverage points are points that appear to be outlying in the regressors. Methods such as studentized deleted residual and Mahalanobis distance are used to detect the presence of outliers and leverage point respectively (Oyeyemi et al., 2016). Cook and Dffits (1977), are often used to determine if either the outliers or leverage points influences the regression coefficients. Robust regression estimator is commonly used to circumvent the problem of outlier.

II. METHODOLOGY

The linear regression model considered in this study is of two independent variables of the form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1,2 \ldots, n \quad 1$$

Where, y_i is a dependent variable and x_1i and x_2i are independent variables and ε_i is a disturbance (error term) which is identically and normally distributed with zero mean and positive variance. The two explanatory variables in the model are normal variates generated to exhibit some degree of multicollinearity (ρ= 0, 0.2, 0.4, 0.6, 0.8 and 0.99) and some percentage of outlier (δ=1, 2, 3, 4, 5 and 6).

The study therefore, examined and compared the performances of two methods of parameter estimation of multiple linear regression model namely; Least Absolute Deviation (LAD) and Weighted Least Square (WLS) with a view to identify the best method(s) under the conditions stated earlier

Least Absolute Deviation

This estimator obtains a higher efficiency through minimizing the sum of the absolute errors:

$$\min \sum_{i=1}^{n} |\varepsilon_i| \quad 2$$

By considering the objective function:

$$f(\beta) = \left||Y - \beta X|\right|$$

$$f(\beta) = \sum_{i=1}^{n} \left|Y_i - \sum_{j=1}^{m} \beta_j X_{ij}\right|$$

Differentiating this objective function is a problem, since it involves absolute values However, the absolute value function:g(z)=|z|
is differentiable everywhere except at one point: z = 0. Furthermore, by applying the following simple formula for the derivative, where it exists

$$g'(z) = \frac{z}{|z|}$$

Using this formula to differentiate f with respect to each parameter, and setting the derivatives to zero, gives following equations for critical point

$$\frac{\partial f}{\partial \beta_r} = \sum_{i=1}^{n} \frac{Y_i - \sum_{j=1}^{m} \beta_j X_{ij}}{\left|Y_i - \sum_{j=1}^{m} \beta_j X_{ij}\right|} (-X_{ir}) = 0$$

where r =1, 2, …., m

rewrite as:

$$\sum_{i=1}^{n} \frac{\beta_j X_{ij}}{\varepsilon_i} = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\beta_j X_{ir} X_{ij}}{\varepsilon_i} \quad 3$$

Let (w) denote the diagonal matrix, (Dasgupta and Mishra, 2004), where:

$$w_{ij} = \frac{1}{|\varepsilon_i|} for i = j$$

$$w_{ij} = 0 \;\; for i \neq j$$

The equation (3) in matrix notation as follows:

$$(X'WY) = X'WY\beta \quad 4$$

This equation can't be solved for *x*. But let rearrange this system of equations by pre-multiplying both sides by$(X'WX)^{-1}$

$$\hat{\beta} = (X'WX)^{-1}X'WX$$

This formula suggests an iterative scheme that hopefully converges to a solution. Indeed, by initializing $\beta^{(0)}$ arbitrarily and then use the above formula to successively compute new approximations. By let $\beta^{(k)}$ denote the approximation at the *kth* iteration, then the update formula can be expressed as:

$$\hat{\beta}_{LAD}^{(k)} = (X'WX)^{-1}X'WX \quad 5$$

Mean Square Error for LAD Model

$$(MSE)_{LAD} = (\sigma^2)_{LAD} \quad 6$$

$$= SST - SSR = Y'WY - (\hat{\beta})'X'WY \quad 7$$

Mean Square Error for LAD Estimator

$$MSE(\hat{\beta})_{LAD} = (\sigma^2)_{LAD} tr(X'WX)^{-1} \quad 8$$

Mean Absolute Error for LAD

$$MAE = \frac{\sum_{i=1}^{n}|\varepsilon_i|}{n} \quad 9$$

Absolute Bias for LAD

$$Absolute\ (Bias)_{LAD} = \left|E(\hat{\beta})_{LAD} - \beta_{LAD}\right| \quad 10$$

$$Mean\ Absolute\ ((Bias)_{LAD} = \frac{\sum_{i=1}^{n}|\bar{y}-\mu|}{n} \quad 11$$

Weighted Least Square

When applying ordinary least squares to estimate linear regression, (naturally) minimize the mean squared error:

$$MSE(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 \quad 12$$

The solution is of course

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

Could instead minimize the weighted mean squared error,

$$WMSE(\beta, w_1, \dots, w_n) = \frac{1}{n}\sum_{i=1}^{n} w_i(y_i - x_i\beta)^2$$

This includes ordinary least squares as the special case where all the weights $w_i = 1$. Solve it by the same kind of linear algebra to solve the ordinary linear least squares problem. By writing **w** for the matrix with the w_i on the diagonal and zeroes everywhere else, then,

$$WMSE = n^{-1}(Y - X)'W(Y - X\beta)$$

$$= \frac{1}{n}(Y'WY - Y'WX\beta - \beta'X'WY + \beta'X'WX\beta)$$

Differentiating with respect to β', gives as the gradient

$$\nabla_\beta WMSE = \frac{2}{n}(-X'WY + X'WX\beta)$$

Setting this to zero at the optimum and solving,

$$\hat{\beta}_{WLS} = (X'WX)^{-1}X'WX \quad 13$$

Absolute Bias for WLS

$$Absolute\ (Bias)_{WLS} = \left|E(\hat{\beta})_{WLS} - \beta_{WLS}\right| \quad 14$$

$$Mean\ Absolute\ ((Bias)_{WLS} = \frac{\sum_{i=1}^{n}|\bar{y}-\mu|}{n} \quad 15$$

Algorithms for Model Specification

The model considered in the simulation is equation (1.1),

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i, i = 1,2 \ldots, n$$

where, y is dependent variable, $x_1\ and\ x_2$are two independent variables, β_j , $j = 0,1,2$ are parameters of the regression and ε_i is random error.

The explanatory variables used in this study were generated with specified inter-correlations (level of multicollinearity) as follows;

$$z_1 = \frac{x_1-\mu_1}{\sigma_1} \Rightarrow x_1 = \mu_1 + \sigma_1 z_1$$

$$z_2 = \frac{x_2-\mu_2}{\sigma_2} \Rightarrow x_2 = \mu_2 + \sigma_2 z_2.$$

$$\rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \Rightarrow \sigma_{12} = \rho\sigma_1\sigma_2$$

But to introduce correlation into the two independent variables, we assumed that the x_2 is linearly dependent on x_1 in the following form;

$$x_2 = \propto_0 + \propto_1 x_1 + e_i$$

$$\propto_1 = \frac{\sigma_{12}}{\sigma_1^2} \Rightarrow \sigma_{12} = \propto_1 \sigma_1^2 \Rightarrow \sigma_2 = \frac{\propto_1\sigma_1}{\rho}$$

Therefore,

$$x_2 = \mu_2 + \frac{\propto_1\sigma_1}{\rho} z_2$$

16

Data were simulated for both exogenous variables and error terms from normal distribution with mean zero and variance one i.e; $z_{1i} \sim N(0,1), z_{2i} \sim N(0,1) and\ \varepsilon_i \sim N(0,1), i = 1,2, \ldots, 1000\ (iteration)$.

The values of explanatory variables were obtained from relations (16) and (17) at different levels of multicollinearity.

Evaluation, Comparison and Preference of Estimators

At each scenario of specification, inter-correlation between the two exogenous variables (multicollinearity level), percentage of outlier and sample size, the estimators were examined and compared using the finite sampling properties of estimators which are absolute bias (AB), mean absolute error (MAE) and mean squared error (MSE) criteria. The estimator with minimum criteria under different scenario of simulation were taken as the best.

III. ANALYSIS AND RESULTS

The effect of different levels of multicollinearity (ρ = 0, 0.2, 0.4, 0.6, 0.8, 0.99) and percentage of Outlier (δ=1,2,3,4,5 and 6) at the sample size of 10, 20 and 50 which represent small, moderate and large sample sizes respectively on the simulated data from the multiple linear regression in 1. The simulation study was carried out with 1000 iteration on each case in R statistical software. For each iteration, the values of the criteria for the assessment (MSE, MAE and AB) were computed and their average values were recorded according to sample sizes as shown in table 1-6.

1

Table 1. Results of Performance of Estimators for Different Levels of Multicollinearity (ρ) when Sample Size is 10 (Small)

(ρ)	LAD	WLS

	MSE	MAE	AB	MSE	MAE	AB
0	1.040E-02	3.100E-02	1.048E-03	0.706	0.658	0.004
0.2	1.723E-02	3.132E-02	1.246E-03	0.718	0.656	0.005
0.4	3.150E-02	3.319E-02	1.359E-03	0.721	0.661	0.005
0.6	3.159E-02	3.981E-02	1.978E-03	0.725	0.662	0.006
0.8	3.411E-02	4.495E-02	2.342E-03	0.727	0.664	0.007
0.99	6.521E-02	7.119E-02	3.725E-03	0.728	0.665	0.012

It was observed that LAD estimator is better than WLS, because it has the minimum values of the three criteria used for the assessment. Furthermore, the performances of the two methods decreased when the level of multicollinearity was increased.

Table 2. Results of Performance of Estimators for Different Levels of Multicollinearity when Sample Size is 20 (Medium)

(ρ)	LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB
0	2.214E-02	4.173E-01	2.578E-04	0.0849	0.729	0.0011
0.2	4.167E-02	4.190E-01	2.944E-04	0.0851	0.730	0.0012
0.4	4.256E-02	4.462E-01	5.528E-04	0.0864	0.734	0.0013
0.6	4.976E-02	4.978E-01	5.928E-04	0.0866	0.735	0.0013
0.8	4.763E-02	5.877E-01	7.856E-04	0.0867	0.737	0.0014
0.99	8.100E-02	6.429E-01	9.432E-04	0.0871	0.740	0.0015

From the table 2 above, LAD was the best estimators while WLS has the least performance among the two at all levels of multicollinearity and sample size of 20. However, the performances of the two estimators decreased when the level of multicollinearity was increased.

Table 3 Results of Performance of Estimators at Different Levels of Multicollinearity when Sample Size is 50 (Large)

(ρ)	LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB
0	2.416E-01	1.052E-02	1.984E-05	0.941	0.673	1.000E-04
0.2	4.829E-01	4.793E-02	2.046E-05	0.943	0.674	1.200E-04
0.4	4.893E-01	4.822E-02	2.536E-05	0.945	0.774	1.170E-04
0.6	4.993E-01	4.999E-02	2.945E-05	0.947	0.774	1.172E-04
0.8	5.399E-01	5.093E-02	3.616E-05	0.947	0.775	1.227E-04
0.99	9.379E-01	6.672E-02	8.877E-05	0.949	0.776	1.580E-04

The average values of MSE, MAE and AB recorded in table 3 show that LAD was the best estimators while WLS has the least performance among the two at all levels of multicollinearity. The gaps between their performances increases relatively especially on the basis MAE to that of sample size of 20. However, the performances of the two estimators decreased when the level of multicollinearity was increased.

Table 4. Results of Performance of Estimators for Different percentage of Outlier (δ) when Sample Size is 10 (Small)

(δ)	LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB
10%	4.365E-01	1.095E-01	1.165E-03	0.682	0.651	0.051

20%	5.026E-01	1.348E-01	1.546E-03	0.693	0.651	0.059
30%	6.664E-01	1.558E-01	1.638E-03	0.702	0.655	0.063
40%	6.769E-01	1.757E-01	1.993E-03	0.714	0.664	7.21E-02
50%	6.769E-01	2.757E-01	2.993E-03	0.719	0.664	7.29E-02
60%	7.204E-01	3.081E-01	3.574E-03	0.721	0.665	8.53E-02

LAD was the best estimators while WLS has the least performance among the two at all percentage of outlier. WLS compute with LAD as the percentage of outlier was increased from 30% and above especially on the basis MSE.

Table 5 Results of Performance of Estimators at Different Levels of Outlier (δ) when Sample Size is 20 (Medium)

(δ)	LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB
5%	9.992E-02	2.016E-01	2.644E-05	0.842	0.729	0.0003
10%	1.061E-01	2.114E-01	3.156E-05	0.859	0.729	0.0003
15%	1.061E-01	2.114E-01	3.156E-05	0.859	0.733	0.0004
20%	1.130E-01	2.226E-01	4.589E-05	0.862	0.739	0.0005
25%	1.130E-01	2.526E-01	5.889E-05	0.868	0.759	0.0006
30%	1.130E-01	2.826E-01	7.989E-05	0.876	0.769	0.0008

From the table above, LAD was the best estimators while WLS has the least performance among the two at all percentage of outlier. However, the performances of the two estimators decreased when the percentage of outlier was decreased.

Table 6 Results of Performance of Estimators for Different Levels of Outlier (δ) when Sample Size is 50 (Large)

(δ)	LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB
2%	1.154E-01	2.482E-02	1.373E-05	0.940	0.772	3.16E-04
4%	1.191E-01	2.548E-02	1.818E-05	0.940	0.773	3.86E-04
6%	1.206E-01	2.836E-01	4.763E-05	0.946	0.775	3.86E-04
8%	1.229E-01	2.998E-01	5.324E-05	0.948	0.777	5.03E-04
10%	1.249E-01	3.108E-01	5.424E-05	0.949	0.778	6.13E-04
12%	1.275E-01	3.437E-01	6.587E-05	0.953	0.779	6.91E-04

LAD was the best estimators while WLS has the least performance among the two at all percentages of outlier. The gaps between their performances increases relatively most especially for MSE for sample size of 50. However, the performances of the two estimators decreased when the level of outlier was decreased.

IV. CONCLUSION

This study has revealed that the LAD was the best when there is no multicollinearity or outlier. When some levels of multicollinearity and outlier were increased in the data LAD still maintained the best. However, the performances of the two estimators increased when the level of multicollinearity and outlier was decreased. Furthermore, the performance of WLS estimator decreases with increases in sample size especially for sample size of 10.

REFERENCES

1. Adnan,N; Ahmad,M. H; and Adnan,R. (2006), " Comparative Study on some methods for

Handling Multicollinearity problems" MATEMATIKA, Vol. 22,No. 2, pp 109-119.

2. Ajiboye S. Adegoke E. A. Kayode A. and Adewale F. (2016). A Comparative Study of Some Robust Ridge and Liu Estimators, Science World Journal 11 (4): 16 – 20.
3. Akeyede, I., Ailobhio, D.T. and Ayoo P.V. (2017). Relative Efficiency of Ridge Regression and Ordinary Least Square Estimators on Linear Regression Models at Different Level of Multicollinearity. FULAFIA Journal of Science and Technology Vol 3, (2): 77 – 81.
4. Alkhamisi, M.A. and G. Shukur, (2008). Developing Ridge Parameters for SUR Model. Commun. Stat. Theory Methods, 37: 544 – 564.
5. Alkhamisi, M., G. Khalaf and G. Shukur, (2006). Some Modifications for Choosing Ridge Parameters.Commun. Stat., Theory Methods, 35: 2005 – 2020.
6. Barnett V., and Lewis, T.(1984), Outliers in Statistical Data, John Wiley & Sons, Chichester.
7. Cook and Dffits, (1977). Detection of influential observations in linear regression. Technometrics.
8. Grubbs, F.E. (1969), Procedures for Detecting Outlying Observations in Samples, Technometrics, 11, 1-21
9. Gumbel, E.J. (1960), Discussion on Rejection of Outliers by Anscombe, F. J. Technometrics , 2, 165-166.
10. Frisch, R (1934): Statistical Confluence Analysis by Means of Complete Regression Systems. Oslo: University Economics Institute.
11. Habshah, M. , Sohel R., A.H.M. and Rahmatullah I. (2014). Two-Step Robust Estimator in Heteroscedastic Regression Model in the Presence of Outliers.
12. Hoerl, A. E. and Kennard, R.W. (1970). Ridge Regression Biased Estimation for Nonorthogonal Problems. Technometrics, 12, 55-67.
13. Khalaf, G. and G. Shukur, (2005). Choosing Ridge Parameters for regression problem. Commun. Stat. Theory Methods, 34: 1177 – 1182.
14. Muniz, G. and B.G. Kibria, (2009). On Some Ridge Regression Estimators: An Empirical Comparisons. Commun. Stat. Simul. Comput., 38: 621 – 630.
15. Oyeyemi, G.M., Bukoye, A. and Akeyede, I. (2016) Comparison of Outlier Detection Procedures in Multiple Linear Regression, American Journal of Mathematics and Statistics. 5(1): 37-41.
16. Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proceedings of the Third Berkeley symposium on mathematical statistics and probability, 1, 197– 20.
17. Valentine, T. J (1969): A Note on Multicollinearity. Australian Economic Papers. 8 pp.99 – 105.
18. Wilcox, R. R. (2010). Fundamentals of Modern Statistical Methods", 2nd. Edition, Springer, London..