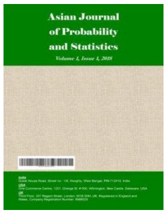
Asian Journal of Probability and Statistics
Volume 1, Issue 1, 2018

Asian Journal of Probability and Statistics

2(2): 1-8, 2018; Article no.AJPAS.45509

A Comparative Analysis on Some Estimators of Parameters of Linear Regression Models in Presence of Multicollinearity

Warha, Abdulhamid Audu[1*], Yusuf Abbakar Muhammad[2] and Akeyede, Imam[3]

[1]Department of Mathematics and Statistics, Ramat Polytechnic Maiduguri, Nigeria.
[2]Department of Mathematical Sciences, University of Maiduguri, Nigeria.
[3]Department of Mathematics, Federal University Lafia, P.M.B. 146 Lafia, Nigeria.

Authors' contributions

This work was carried out in collaboration between all authors. Author WAA designed the study, performed the statistical analysis, wrote the protocol, and wrote the first draft of the manuscript. Author YAM managed the analyses of the study. Author AI managed the literature searches. All authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2018/v2i228773
Editor(s):
(1) Dr. H. K. Nigam, Associate Professor, Department of Mathematics, Faculty of Engineering and Technology, Mody Institute of Technology and Science (Deemed University), Laxmangarh, India.
Reviewers:
(1) Zlatin Zlatev, Trakia University, Bulgaria.
(2) Kunio Takezawa, Institute for Agro-Environmental Sciences, NARO (NIAES), Japan.
Complete Peer review History: http://www.sciencedomain.org/review-history/27831

Received: 20 September 2018
Accepted: 24 November 2018
Published: 18 December 2018

Original Research Article

Abstract

Linear regression is the measure of relationship between two or more variables known as dependent and independent variables. Classical least squares method for estimating regression models consist of minimising the sum of the squared residuals. Among the assumptions of Ordinary least squares method (OLS) is that there is no correlations (multicollinearity) between the independent variables. Violation of this assumptions arises most often in regression analysis and can lead to inefficiency of the least square method. This study, 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 in the explanatory variables. Simulation techniques were conducted using R Statistical software, to investigate the performance of the two estimators under violation of assumptions of lack of multicollinearity. 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 and sample size. The results showed that, LAD was the best at

*Corresponding author: E-mail: abdulhamidaudu411@gmail.com;

different levels of multicollinearity and was recommended as alternative to OLS under this condition. The performances of the two estimators decreased when the levels of multicollinearity was increased.

Keywords: LAD; WLS; multicollinearity; regression; simulation.

1 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), X_1, X_2, …, X_p. 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) [1]. In the situation where the assumptions of the linear regression are not valid, many estimation methods have been proposed; Stein Estimator by Stein [2], Liu Estimator by Liu [3], and Ridge Estimator proposed by Hoerl and Kennard [4], 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 a correlation in the explanatory variables which are referred to as multicollinearity 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.

1.1 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 [5]. 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 [6].

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 [7], as noted by Valentine [8], 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 [9], Khalaf and Shukur [10] Alkhamisi et al. [11], Alkhamisi and Shukur [12]), 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.

2 Methodology

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

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

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

For the simulation study, the parameters of the model above were fixed $as\, \beta_j = 1,\ j = 0,1,2$. The Multicollinearity levels (ρ) are 0, 0.2, 0.4, 0.6, 0.8 and 0.99, which indicate the strength of multicollinearity from the lowest to highest levels are introduce into the explanatory variables.

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 minimising the sum of the absolute errors:

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

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 points

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

where r =1, 2, …., m

rewrite as:

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

Let (w) denote the diagonal matrix, [13], where:

$$w_{ij} = \frac{1}{|\varepsilon_i|} for\ i = j$$
$$w_{ij} = 0\ fori \neq j$$

The equation (3) in matrix notation as follows:

$$(X'WY) = X'WY\beta \tag{4}$$

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 initialising $\beta^{(0)}$ arbitrarily and then use the above formula to successively compute new approximations. By let $\beta^{(k)}$ denote the approximation at the *k^th^* iteration, then the update formula can be expressed as:

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

Mean Square Error for LAD Model:

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

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

Mean Square Error for LAD Estimator:

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

Mean Absolute Error for LAD:

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

Absolute Bias for LAD:

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

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

Weighted Least Square:

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

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

The solution is of course

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

Could instead minimise 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 $w_i = 1$. 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,

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

Absolute Bias for WLS:

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

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

Algorithms for Model Specification:

The model considered in the simulation is equation (1),

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

where, y is dependent variable, $x_1\ and\ x_2$are two independent variables, $\beta_j\ , j = 0,1,2$ 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;

$$z_1 = \frac{x_1 - \mu_1}{\sigma_1} \Rightarrow x_1 = \mu_1 + \sigma_1 z_1 \tag{16}$$

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

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

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,

$$x_2 = \mu_2 + \frac{\propto_1 \sigma_1}{\rho} z_2 \qquad (17)$$

Data were simulated for both exogenous variables and error terms from normal distribution with mean zero and variance one i.e;

$$z_{1i} \sim N(0,1), z_{2i} \sim N(0,1)\ and\ \varepsilon_i \sim N(0,1), i = 1,2,\dots,1000\ (iteration).$$

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) 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 was taken as the best.

3 Analysis and Results

The effect of different levels of multicollinearity (ρ = 0, 0.2, 0.4, 0.6, 0.8, 0.99) 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 Tables 1 – 3.

Table 1. Results of performance of estimators for different levels of multicollinearity (ρ) when sample size is 10 (Small)

(ρ)	OLS			LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB	MSE	MAE	AB
0	1.140E-03	2.300E-02	0.048E-03	1.040E-02	3.100E-02	1.048E-03	0.706	0.658	0.004
0.2	1.823E-03	2.532E-02	0.146E-03	1.723E-02	3.132E-02	1.246E-03	0.718	0.656	0.005
0.4	2.150E-03	2.719E-02	0.409E-03	3.150E-02	3.319E-02	1.359E-03	0.721	0.661	0.005
0.6	2.859E-02	2.901E-02	0.818E-03	3.159E-02	2.181E-02	1.978E-03	0.725	0.662	0.006
0.8	3.211E-02	3.295E-02	1.101E-03	3.411E-02	3.095E-02	2.342E-03	0.727	0.664	0.007
0.99	5.821E-02	6.019E-02	2.705E-03	6.521E-02	5.119E-02	3.725E-03	0.728	0.665	0.012

It was observed that OLS was the best estimators because it has the minimum values of the three criterial used for the assessment followed by LAD while WLS has the least performance among the three estimators. However the LAD compete with OLS as the level of multicollinearity was increased and even performed better than others from multicollinearity above 0.4 especially on the basis MAE. Furthermore, the performances of the three methods decreased when the level of multicollinearity was increased, i.e with decrease in their critical values, due to the inecrease in the strength of correlation between the two explanatory variables.

Table 2. Results of performance of estimators for different levels of multicollinearity when sample size is 20 (Medium)

(ρ)	OLS			LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB	MSE	MAE	AB
0	2.814E-02	4.723E-01	2.878E-04	2.214E-02	4.173E-01	2.578E-04	0.0849	0.729	0.0011
0.2	4.967E-02	4.990E-01	3.104E-04	4.167E-02	4.190E-01	2.944E-04	0.0851	0.730	0.0012
0.4	5.056E-02	5.002E-01	5.689E-04	4.256E-02	4.462E-01	5.528E-04	0.0864	0.734	0.0013
0.6	5.176E-02	5.178E-01	6.152E-04	4.976E-02	4.978E-01	5.928E-04	0.0866	0.735	0.0013
0.8	5.863E-02	6.007E-01	8.001E-04	4.763E-02	5.877E-01	7.856E-04	0.0867	0.737	0.0014
0.99	8.902E-02	6.929E-01	9.930E-04	8.100E-02	6.429E-01	9.432E-04	0.0871	0.740	0.0015

From the Table 2, LAD was the best estimators follow by OLS while WLS has the least performance among the three estimators at all levels of multicollinearity and sample size of 20. However, the performances of the three 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)	OLS			LAD			WLS		
	MSE	MAE	AB	MSE	MAE	AB	MSE	MAE	AB
0	2.916E-01	1.752E-02	2.012E-05	2.416E-01	1.052E-02	1.984E-05	0.941	0.673	1.000E-04
0.2	5.019E-01	4.993E-02	2.546E-05	4.829E-01	4.793E-02	2.046E-05	0.943	0.674	1.200E-04
0.4	5.421E-01	5.122E-02	2.836E-05	4.893E-01	4.822E-02	2.536E-05	0.945	0.774	1.170E-04
0.6	5.845E-01	5.209E-02	3.100E-05	4.993E-01	4.999E-02	2.945E-05	0.947	0.774	1.172E-04
0.8	5.999E-01	5.593E-02	3.821E-05	5.399E-01	5.093E-02	3.616E-05	0.947	0.775	1.227E-04
0.99	9.819E-01	6.801E-02	9.123E-05	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 follow by OLS while WLS has the least performance among the three 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.

4 Conclusion

This study has revealed that the LAD was the best when the sample size were increased especially on the basis sample size 20 medium and 50 large while OLS was the best for small sample size (10 small), when there is no multicollinearity. When some levels of multicollinearity were increased in the data LAD still maintained the best for sample size of 20 medium and 50 large. However, the performances of the three estimators increased when the level of multicollinearity was decreased. Furthermore, WLS has the best when there is multicollinearity at sample size of 20 in some cases.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Ajiboye S, Adegoke EA, Kayode A, Adewale F. A comparative study of some robust ridge and liu estimators. Science World Journal. 2016;11(4):16-20.

[2] Stein C. 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. 1956;1: 197-20.

[3] Liu K. A new class of biased estimate in linear regression. Communication in Statistics. 1993;22(2):393-402.

[4] Hoerl AE, Kennard RW. Ridge regression biased estimation for nonorthogonal problems. Technometrics. 1970;12:55-67.

[5] Adnan N, Ahmad MH, Adnan R. Comparative Study on some methods for handling multicollinearity problems. MATEMATIKA. 2006;22(2):109-119.

[6] Akeyede I, Ailobhio DT, Ayoo PV. 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. 2017;3(2):77–81.

[7] Frisch R Statistical confluence analysis by means of complete regression systems. Oslo: University Economics Institute; 1934.

[8] Valentine TJ. A note on multicollinearity. Australian Economic Papers. 1969;8:99 –105.

[9] Muniz G, Kibria BG. On some ridge regression estimators: An empirical comparisons. Commun. Stat. Simul. Comput. 2009;38:621-630.

[10] Khalaf G, Shukur G. Choosing ridge parameters for regression problem. Commun. Stat. Theory Methods. 2005;34:1177–1182.

[11] Alkhamisi M, Khalaf G, Shukur G. Some modifications for choosing ridge parameters. Commun. Stat., Theory Methods. 2006;35:2005–2020.

[12] Alkhamisi MA, Shukur G. Developing ridge parameters for SUR model. Commun. Stat. Theory Methods. 2008;37:544–564.

[13] DasGupta M, Mishra SK. Least absolute deviation estimation of linear econometric models: A literature review; 2004. Available:http://mpra.ub.uni-muenchen.de/1781.

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sciencedomain.org/review-history/27831