BAREKENG: Journal of Mathematics and Its Application
September 2022 Volume 16 Issue 3 Page 779–786
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol16iss3pp779-786

RIDGE LEAST ABSOLUTE DEVIATION PERFORMANCE IN ADDRESSING MULTICOLLINEARITY AND DIFFERENT LEVELS OF OUTLIER SIMULTANEOUSLY

Netti Herawati[1*], Subian Saidi[2], Dorrah Azis[3]

[1,2,3]*Department of Mathematics, Faculty of Matematics and Natural Sciences, University of Lampung*
Prof. Dr. S. Brojonegoro St. No.1, Bandar Lampung, Lampung, 35141, Indonesia

Corresponding author's e-mail: [1]* *netti.herawati@fmipa.unila.ac.id*

Abstract. *If there is multicollinearity and outliers in the data, the inference about parameter estimation in the LS method will deviate due to the inefficiency of this method in estimating. To overcome these two problems simultaneously can be done using robust regression, one of which is the ridge least absolute deviation method. This study aims to evaluate the performance of the ridge least absolute deviation method in surmounting multicollinearity in diverse sample sizes and percentage of outliers using simulation data. The Monte Carlo study was designed in a multiple regression model with multicollinearity (ρ=0.99) between variables x_1 and x_2 and outliers of 10%, 20%, and 30% on response variables with different sample sizes (n = 25, 50,75,100,200; β_0=0, and β=1 otherwise). The existence of multicollinearity in the data is done by calculating the correlation value between the independent variables and the VIF value. Outlier detection is done by using boxplot. Parameter estimation was carried out using the RLAD and LS methods. Furthermore, a comparison of the MSE values of the two methods is carried out to see which method is better at overcoming multicollinearity and outliers. The results showed that RLAD had a lower MSE than LS. This means that RLAD is more precise in estimating the regression coefficients for each sample size and the various outlier levels studied.*

Keywords: *multicollinearity, outliers, RLAD, LS, MSE.*

Article info:

Submitted: 7th March 2022 *Accepted: 8th July 2022*

How to cite this article:
Herawati, N. Saidi, S and Azis, D. "RIDGE LEAST ABSOLUTE DEVIATION PERFORMANCE IN ADDRESSING DIFFERENT LEVELS OF OUTLIERS AND MULTICOLLINEARITY SIMULTANEOUSLY", *BAREKENG: J. Math. & App.*, vol. 16, iss. 3, pp. 779-786, September, 2022.

1. INTRODUCTION

As is known in the multiple regression linear model, the presence of multicollinearity and outliers in the data can affect the conclusion if the parameter estimation is carried out using the OLS method. The deviation of the conclusions obtained is caused by the inefficiency of this method if the assumptions of multicollinearity [1-4]. Using robust regression to cope with the correlation between independent variables can be done. There are several types of robust methods, one of which is robust ridge regression. The advantage of using the robust ridge regression method is that the standard error can be reduced and a more accurate estimation of the regression coefficient can be obtained [5-10].

In addition, deviations in conclusions can also be found when there is data that deviates from the average which is called an outlier. Outliers can lead to non-fulfillment of the assumption of normality and homogeneous error [11]. We need a method that can solve the problems of outliers and multicollinearity simultaneously.

Regression Ridge Least Absolute Deviation (RLAD) is an alternative method that we chose to handle multicollinearity and outliers at the same time, besides other robust methods that are widely available. This method is a combination of the robust ridge regression method and the least absolute deviation method. The RLAD estimator that results will be stable and resistant to outliers [12]. However, there has been no comprehensive research using this method to seek the outcome of this method in overcoming various levels of outliers at various sample sizes. Therefore, in this study, ridge least absolute deviation performance in the multiple regression model with data containing multicollinearity and various levels of outliers at various sample sizes was analyzed and compared with the least squares method using simulation data.

2. RESEARCH METHODS

Least Absolute Deviation (LAD) is a robust regression parameter estimation method that is resistant to the presence of outliers by minimizing the total absolute value of the residual. The least absolute deviation method can be defined as:

$$min \sum_{i=1}^{n} |\varepsilon_i| = min \sum_{i=1}^{n} |Y_i - \boldsymbol{X_i'\beta}_{LAD}| \qquad (1)$$

In the formula, it can be seen that LAD will minimize the absolute value of the residual. This is in contrast to the least squares method, which minimizes the sum of the squares of the residuals. In this way, the effect of outliers will be minimized in the LAD method and will produce more accurate regression coefficient estimator [6].

Because the LAD method does not have an analytical solution to obtain parameter estimates, an iterative approach is needed. The weighted least squares procedure can be used for this. The iterations were performed to obtain a convergent value. The Least Absolute Deviation (LAD) parameter estimator can be solved using the following formula:

$$\widehat{\boldsymbol{\beta}}_{LAD} = (\boldsymbol{X'WX})^{-1}\boldsymbol{X'WY} \qquad (2)$$

where $\boldsymbol{W}$ is a diagonal matrix with diagonal elements w_{ii} with

$$w_{ii} = \begin{cases} \dfrac{1}{|\varepsilon_i|}, & if\ |\varepsilon_i| \neq 0 \\ 1, & if\ |\varepsilon_i| = 0 \end{cases}$$

One of the methods commonly used to solve multicollinearity problems by limiting coefficient estimates is ridge regression, namely by modifying the LS parameter estimator. In this way, ridge regression is able to reduce the variance of the estimator. However, it generates bias. The relatively small constant bias α resulting from the ridge method is added to the main diagonal of the matrix $\boldsymbol{X'X}$ obtained by the LS method to form a new matrix ($\mathbf{X}^{'}\mathbf{X}+ \alpha\mathbf{I}$) [13]. The ridge regression model can be written as:

$$\boldsymbol{Y} = \boldsymbol{X\beta_R} + \boldsymbol{\varepsilon} \qquad (3)$$

where β_R is the ridge regression parameter to be estimated.

The ridge regression estimator requires that $\boldsymbol{\beta}_{\boldsymbol{R}}' \boldsymbol{\beta}_{\boldsymbol{R}} = \boldsymbol{\rho}$ be satisfied to obtain the minimum sum of squares. This can be done using the Lagrange multiplier method to obtain:

$$(\boldsymbol{\beta}_{\boldsymbol{R}}, \alpha) = \boldsymbol{Y}'\boldsymbol{Y} - \boldsymbol{2\beta}_{\boldsymbol{R}}'\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{\beta}_{\boldsymbol{R}}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}_{\boldsymbol{R}} + \alpha(\boldsymbol{\beta}_{\boldsymbol{R}}'\boldsymbol{\beta}_{\boldsymbol{R}} - \boldsymbol{\rho}). \quad (4)$$

If Eq. (4) is derived with respect to $\boldsymbol{\beta}_{\boldsymbol{R}}$ and then equalized to zero, then the following equation will be obtained:

$$\hat{\beta}_R = (\boldsymbol{X}'\boldsymbol{X} + \alpha\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{Y} \quad (5)$$

with I = identity matrix; $0 \leq \alpha \leq 1$; $\hat{\beta}_R$ = ridge parameter vector. Adding bias α to the diagonal matrix $\boldsymbol{X}'\boldsymbol{X}$ in order to get an unbiased estimator coefficient and independent variables X and dependent variable Y should be transformed into a standard variable (standardization) [14].

The combination of LAD and ridge regression will produce the ridge least absolute deviation (RLAD) method. With the ability of LAD to overcome multicollinearity and ridge regression that is able to deal with outliers, the RLAD method will be able to surmount multicollinearity and outliers in the data simultaneously [12]. The parameter estimator of RLAD can be written as:

$$\widehat{\boldsymbol{\beta}}_{RLAD} = (\boldsymbol{X}'\boldsymbol{X} + \alpha^*\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LAD} \quad (6)$$

where $\widehat{\boldsymbol{\beta}}_{LAD}$ = LAD regression estimator with $0 \leq \alpha^* \leq 1$.

There are several ways to select the value of α*. One of the formulas is to use the method introduced by [5, 7, 12] based on the least squares method as follows:

$$\alpha^* = \frac{p\, S_{LAD}^2}{\widehat{\boldsymbol{\beta}}_{LAD}^T \widehat{\boldsymbol{\beta}}_{LAD}} \quad (7)$$

where p = number of independent variables and $S_{LAD}^2 = \frac{(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LAD})^T(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LAD})}{n-p}$.

Mean Square Error (MSE) is one of the most popular and easy to use error measurements. The MSE value is used to measure the accuracy of the estimated value of the regression model, which is expressed in the mean square of the error. Generally, the smaller the MSE, the more accurate the forecast value of a model will be. In addition, in this case, the best method is defined as the method that can fix multicollinearity and outlier problems in unison. The Mean Square Error (MSE) formula to determine the best parameter estimation results of $\hat{\beta}$ is:

$$MSE(\hat{\beta}) = \frac{1}{m}\sum_{j=1}^{m}(\hat{\beta}_j - \beta);\ j = 1,2,\dots,m \quad (8)$$

With $\hat{\beta}_j$ is estimated regression coefficient; β is regression coefficient to be estimate and m is repetition.

Simulated data was used in this study using RStudio 1.2.1335. The true model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ was simulated with different sample sizes (n=20, 40, 60, 100, 200) and $p = 6$ repeated 1000 times. The independent variables $x_{ij} = (1-\rho^2)^{1/2}u_{ij} + \rho u_{ij},\ i = 1,2,\dots,n \quad j = 1,2,\dots,6$ were generated following [15] with, u_{ij} are independent standard normal pseudo-random numbers. Multicollinearity or correlation (ρ=0.99) between variables x_1 and x_2 was designed in the model with 10%, 20%, and 30% outliers in the response variables. β parameters vectors are determined arbitrarily (β_0=0, and β=1 otherwise). Analysis began by testing multicollinearity based on the correlation between the independent variables and the VIF value. Outlier detection was done by using boxplot. Next, calculated the $\widehat{\boldsymbol{\beta}}_{RLAD}$ and $\hat{\beta}_{LS}$. Finally, we compared the MSE values of the two methods.

3. RESULTS AND DISCUSSION

The results of the analysis to ensure the existence of multicollinearity in the data were carried out by calculating the correlation between the independent variables based on VIF values. The results are presented in Table 1.

Tabel 1. VIF values for various sample sizes and 10%, 20% and 30% of outliers for simulated data

	VIF					
n	x_1	x_2	x_3	x_4	x_5	x_6
20	20.8051	24.4644	2.5837	2.2502	1.6855	3.4597
40	40.2781	41.6082	1.7631	2.2338	1.9079	2.1164
60	39.7091	38.7429	1.8391	2.4791	1.9061	1.7895
100	42.9446	43.0693	1.6656	1.7729	1.6564	1.7670
200	28.6906	29.0339	1.8635	1.5891	1.4932	1.7071

As represented in Table 1, the VIF values for the variables x_1 and x_2 are greater than 10. This indicates the presence of multicollinearity. Afterwards, we checked if there were outliers in the data by using a Box plot. We present the box plot for n=20 with 10% outliers as displayed in Figure 1 below.

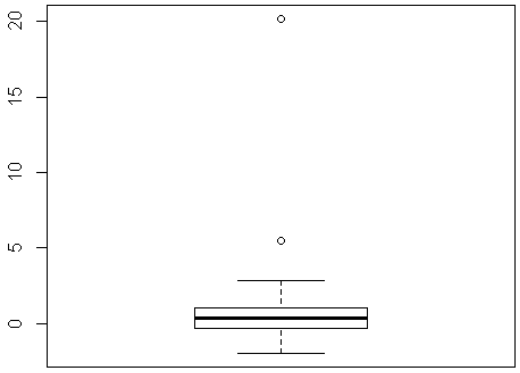

Figure 1. *Boxplot* for n=20 with 10% outliers

Figure 1 shows that outliers were detected using the box plot, which was indicated by data that was far from the average value. The same method was used to detect the presence of outliers for n=20, 40, 60, 100, 200, and outlier levels of 10%, 20%, and 30%. From Table 1 and Figure 1 above, it can be seen that there is multicollinearity indicated by the VIF value and outliers indicated by the data, which is far from the average value in the boxplot. This indicates that a robust method is needed to address the problems. To begin with, we have to reduce the correlation between variables and eliminate outliers by using RLAD. After being analyzed by RLAD, the VIF value of the data was checked again to ensure that there was no longer any correlation between variables. The VIF value generated using RLAD is presented in Table 2.

It can be seen in Table 2 that the value of the correlation between the independent variables is reduced as indicated by the VIF value between the variables. The VIF value of all independent variables becomes less than 10, which indicates that there is no longer a correlation between the independent variables.

Table 2. VIF using RLAD for n=20, 40, 60, 60, 100, 200 with 10%, 20%, and 30% outliers.

		VIF					
n	**outliers**	x_1	x_2	x_3	x_4	x_5	x_6
20	**10%**	0.1946	0.1719	0.3122	0.3171	0.3209	0.3066
	20%	0.1015	0.0895	0.1870	0.1939	0.2070	0.1656
	30%	0.0999	0.0881	0.1790	0.1848	0.1951	0.1624
40	**10%**	0.3127	0.3021	0.5856	0.5940	0.5794	0.6067
	20%	0.1648	0.1589	0.3699	0.3570	0.3627	0.3651
	30%	0.1138	0.1100	0.2626	0.2485	0.2574	0.2532
60	**10%**	0.3951	0.4041	0.7189	0.7182	0.7420	0.7216
	20%	0.1908	0.1965	0.4415	0.4168	0.4455	0.4456
	30%	0.1113	0.1148	0.2747	0.2543	0.2718	0.2773

100	**10%**	0.5781	0.5755	0.8933	0.9230	0.9036	0.9062
	20%	0.3212	0.3199	0.6417	0.6517	0.6492	0.6423
	30%	0.2089	0.2083	0.4558	0.4568	0.4596	0.4530
200	**10%**	1.1339	1.1301	1.1767	1.0860	1.0362	1.1363
	20%	0.6086	0.6033	0.9066	0.8647	0.8365	0.8926
	30%	0.3024	0.2985	0.5978	0.5943	0.5866	0.6010

In addition, the outliers were eliminated automatically. We proceeded to compute the MSE of RLAD and LS. The results are provided in Table 3 and Figure 2.

Tabel 3. MSE of RLAD and LS

Outliers (%)		**n**				
	MSE	**20**	**40**	**60**	**100**	**200**
10	**LS**	91.8476	43.5959	9.6913	7.1296	4.2906
	RLAD	0.7083	0.3524	0.1888	0.1226	0.1199
20	**LS**	101.0431	45.4925	37.3247	11.5228	13.1625
	RLAD	0.9525	0.2516	0.1555	0.1413	0.1153
30	**LS**	102.2833	87.1023	75.4217	47.3645	7.8237
	RLAD	2.0779	0.5166	0.2827	0.2181	0.2043

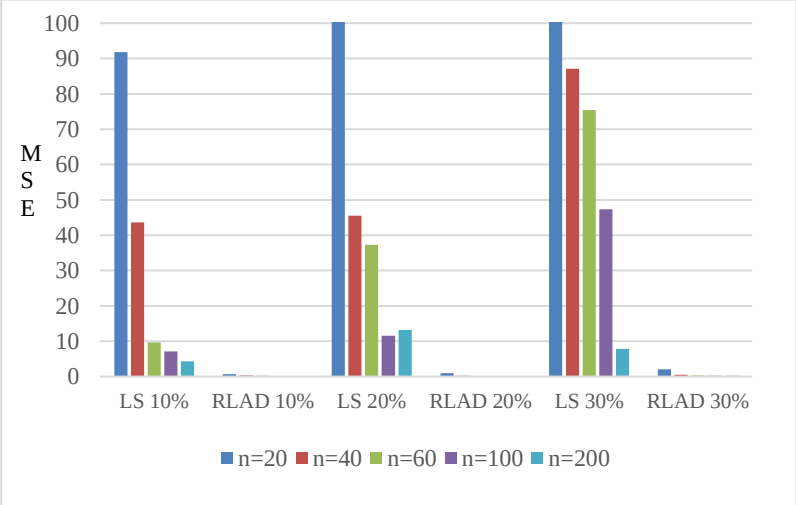

Figure 2. MSE of RLAD and LS

Table 3 and Figure 2 show that RLAD has a smaller MSE than LS for n=20, 40, 60, 100, and 200 for different numbers of outliers in the data. In addition, the sample size seems to also affect the MSE of both methods. Likewise, with the number of outliers in the data. The MSE value decreased with increasing sample size for both methods. However, the MSE for both methods increases as the number of outliers increases.

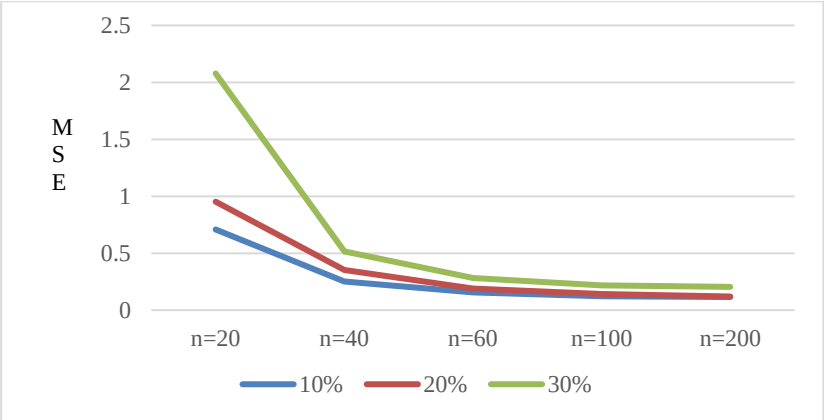

Figure 3. MSE of RLAD for various sample sizes and number of outliers

We present Figure 3 to show MSE RLAD separately from MSE LS to get clearer results. From this figure, it is clear that the MSE of RLAD is higher at small sample sizes. The MSE value decreased with an increasing number of samples. In addition, the MSE value also seems to be influenced by the number of outliers. At a small sample size, the MSE value is high. However, by increasing the sample size, this can be overcome. In addition, Figure 3 shows that the larger the sample size, the outliers will not have much effect on the MSE value.

The results of this study indicate that the RLAD method can overcome multicollinearity and outliers simultaneously compared to the LS method. In addition, when viewed from the MSE value, this study also produces parameter estimates using the RLAD method that are more precise than the LS method. Likewise, if based on the sample size used, it is also found that for small and large sample sizes, the RLAD method can reduce the magnitude of the estimation error value compared to LS. The results of this study are in line with previous results that the robust method in general and the RLAD method in particular can overcome multicollinearity and outliers simultaneously [8-12].

4. CONCLUSIONS

Based on the results of the study, the sample size has an effect on the MSE value of RLAD and LS. The larger the sample size used in the data, the smaller the MSE value for both RLAD and LS, even though the presence of outliers is increasing. Overall, it can be concluded that RLAD has a better performance than LS in overcoming multicollinearity and various outlier levels because it has a smaller MSE value at various sample sizes and levels of outliers studied.

REFERENCES

[1] S. Chatterjee, A.S. Hadi, and B. Price, *Regression Analysis by Example*, 4th ed., vol. 95, no. 452. New Jersey: Jhon Wiley & Sons Inc., 2000.
[2] I. Pardoe, *Applied Regression Modelling*, 3rd ed., New York: Wiley, 2020.
[3] D.C. Montgomery, E.A. Peck and G.G. Vinning., *Introduction to Linear Regression Analysis*. New York: A Wiley Intersection Publication, 2012.
[4] M.H. Kutner et al., *Applied Linear Statistical Models.*, 5th ed. New York: McGraw-Hill, 2005.
[5] D. Gibbons, "A Simulation Study of some Ridge Estimators," *J. Am. Stat. Assoc.*, vol. 76, pp. 131–139, 1981.
[6] M. E. El-Salam, "The Efficiency of Some Robust Ridge Regression for Handling Multicollinearity and Non-Normals Errors Problems.," *Appl. Math. Sci.*, vol. 77, no. 7, pp. 3831 – 3846, 2013.
[7] A. Hoerl and R. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, pp. 55–67, 1970.
[8] N. Herawati, K. Nisa, E. Setiawan, Nusyirwan, and Tiryono, "Regularized Multiple Regression Methods to Deal with Severe Multicollinearity," *Int. J. Stat. Appl.*, vol. 8, no. 4, pp. 167–172, 2018.
[9] E. Setiawan, N. Herawati, K. Nisa, Nusyirwan, and S. Saidi, "Handling Full Multicollinearity and Various Numbers of Outliers using Robust Ridge Regression," *Sci.Int.(Lahore)*, vol. 31, no. 2, pp. 201–204, 2019.
[10] N. Herawati, K. Nisa, and Nusyirwan. " Selecting the Method to Overcome Partial and Full Multicollinearity in Binary Logistic Model," *International Journal of Statistoics and Apllications*, vol. 10, no. 3, pp. 55–59, 2020.

[11] A.M. Leroy and P.J. Rousseeuw, *Robust Regression and Outlier Detection*. Jhon Wiley & Sons, 1987
[12] H. Samkar and O. Alpu, "Ridge Regression Based on Some Robust Estimators.," *J. Mod. Appl. Stat. Methods*, vol. 17, no. 9, pp. 495–501, 2010.
[13] A.E. Hoerl, "Application of ridge analysis to regression problems," *Chem. Eng. Prog.*, vol. 58, pp. 54–59, 1962.
[14] M. El-Dereny and N. I. Rashwan, "Solving Multicollinearity Problem Using Ridge Regression Models.," *Int. J. Contemp. Math. Sci.*, vol. 6, no. 12, pp. 585 – 600, 2011.
[15] G.C. McDonald and D.I. Galarneau, "A Monte Carlo evaluation of some ridge type estimators.," *J. Amer. Stat. Assoc.*, vol. 20, pp. 407–416, 1975.