A Comparison Two Ridge Regression Using LAD method with Simulation

Tamarah Wathib Mohammad\textsuperscript{1*}, Dr.Awatif Rezzoky Al-Dubaicy\textsuperscript{2}

\textsuperscript{1,2} Department Of Mathematics, Education College, Al-Mustansiriyah University, Baghdad ,Iraq

Emails \textsuperscript{1} wathibtamarah@uomustansiriyah.edu.iq
\textsuperscript{2} draraldubaicy@yahoo.com

Abstract. The multicollinearity is the one of the important and contained problems in regression analysis, because its effect on model estimators, the problem is that the independent variables are so closely related that the results of the regression are not clear, the aim of this research is to solve the problem of multicollinearity. one of the solutions get of this problem has deal with, which is the ridge regression of least absolute deviation (LAD) estimators, by adding a proposed a ridge parameter which is considered as contribution to solving the problem of multicollinearity by modify B M Golam Kibria (\(\hat{K}_{ME}\)) then compared it between them. The (\(\hat{K}_{CN}\)) is the best estimator by simulation study and mean square error (MSE) critical.

Keywords. Least Absolute Deviation, Linear Regression, Ridge Regression, Multicollinearity, Errors, Efficiency.

1. Introduction

Multiple linear regressions is an advanced technique that guarantees accurate inference in order to obtain better results of the research by optimizing the use of data for identifying the causal relationships between the phenomena in question. The regression coefficient can be defined as the change that occurs in the response (dependent) variables as any change in the explanatory (independent) variables. The problem lies in the fact that the independent variables are closely related so that the results of the regression are unclear and thus estimating the individual effects of the variables in the regression equation is impossible because the multiple linear relationship can be considered as a case in which two or more independent variables more together and thus it is not possible to specify which of the independent variables caused the noticeable change in the dependent variable\textsuperscript{[13]}. Therefore, in order to solve the problem of linearity, we used the ridge parameter with the Least Absolute Deviation (LAD) method.

Extensive research has been done on the performance of ridge regression and the LAD method, separately. Both of methods are robust, but each of them suits a different type of the problems. An early attempt to take the two methods together in one estimation procedure is from Pfaffenberger and Dielman (1989)\textsuperscript{[15]}, when they tried to find the ridge biasing parameter k through the LAD estimations of each of the parameters and
the variance of the error terms. In other words, they employed the same biasing parameters of Hoerl et al. (1975)\(^6\), but instead of using the Ordinary Least Square (OLS) estimate the parameters and the variance of error terms in the formula, they used the LAD estimations of them.

The (LAD) method provides a strong alternative to Ordinary Least Square (OLS), especially when the data follows abnormal distribution and is subject to outliers. (LAD) estimates are not influenced too much by extreme values, relative to (OLS) estimates. On the other hand, although (LAD) was proposed as an alternative to (OLS) regression method, it is less used and therefore it can be considered a nontraditional technique \(^3\). Ridge regression was used because it is one of the solutions to solve the problem of linear multiplicity. A. E. Hoerl (1962)\(^5\) he was the first to suggest the ridge parameter to control the inflation and general instability. Hoerl and Kennard (1970)\(^8\) the \(K_i\) value reduces MSE(\(\hat{\beta}(K)\)). Hocking et al. (1976)\(^7\) he states that for the known optimum \(K_i\), generating estimator outperforms all other estimates within the class of biased estimators they considered. However, the optimum value of \(K\) is completely dependent on unknown parameters and must be estimated. Hoerl and Kennard (1970)\(^8\), suggested to replace unknown parameter by their corresponding unbiased estimators, Hoerl, Kennard and Baldwin (1975)\(^6\) \((\hat{K}_{HKB}\) or HKB) proposed a different estimator of \(K\) by selecting the harmonic mean. From the Bayesians view point, suggested Lawless and Wang (1976)\(^1\) \((\hat{K}_{LW}\) or LW), Hocking, Speed and Lynn (1976)\(^7\) \((\hat{K}_{HSL}\) or HSL) proposed estimator for \(K\), B M Golam Kibria (2003)\(^9\) he proposed estimating the ridge parameter using the arithmetic mean and geometric mean and median. In this research, took three ridge parameters above and proposed a new ridge parameter, then compared the estimator we proposed with the estimator B M Golam Kibria \((\hat{K}_{MED}\) as the proposed method for the ridge parameter represents the ridge parameter with subtracting the condition number value which is regarded as a ratio of the largest to the smallest singular value\(^1\) then simulation study used to compare the two estimators and find out which of them is the best through mean square error (MSE), as it is the least (MSE).

The paper has been organized as following: In section (2), the main idea for (OLS) and (OLS -ridge) has been discussed. In section (3), the main idea for (LAD) and (LAD -ridge) has been discussed also, ridge parameter has been reviewed. In section (4), proposed ridge parameter. In section (5), the ridge parameters efficiency has been verified and using the simulation method to comparison between these parameters, the results have been analyzed.

2. Ordinary Least Square Estimator (OLS)

The (OLS) method is one of the best and oldest and it is one of the most common unbiased estimation methods are used to estimate the parameters of linear regression model:-

\[ y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i \quad i=1,2,\ldots,n \quad (1) \]

Where

- \(n\): Observations number.
- \(p\): Independent variables number.
- \(x\): The independent variables.
- \(y\): The independent variables.
- \(\epsilon\): Error model.
- \(\beta\): The regression coefficients.

and the (OLS) method works to minimization the sum of squares deviations as which is unbiased estimators of the \(\beta\) as \(E(\hat{\beta}_{OLS})=\beta\), has minimal variability as \(V(\hat{\beta}_{OLS})=\sigma^2(X'X)^{-1}\) which is attains the L.b. of Rao-Gramere inq., so it is best Linear unbiased estimators (BLUE) for \(\beta\) as:

\[ \hat{\epsilon} = (Y - X\hat{\beta})'(Y - X\hat{\beta}) = \]

\[ Y'Y - \beta'XY - Y'X\beta + \beta'XX\beta \]

\[ Y'Y - 2\beta'XY + \beta'XX\beta \]

then by taking the derivation for last equation with respect to \(\beta\) and equal to zero get:-

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

As many scientists worked to find other solutions to processing the problem of linearity and one of these methods was that they proposed ridge parameters\(^{13}\).
2.1 Ordinary Ridge Regression Estimators

Different methods have been proposed to deal with common linear data by modifying the (OLS) method, by introduction some bias on the estimations of the regression parameters. The most common methods is the ridge regression method, it is used to solve the problem of linear multiplicity, where the ridge regression estimates depend on the parameter (K) called ridge parameter or biasing constant, its value $K \geq 0$ chosen by the researcher according to some suitable Criteria set by Hoerl and Kennard [8] then the ridge regression estimator ($b_{RR}$) as follows:

$$b_{RR} = (X'X + KI)^{-1}X'Y$$

(3)

Where the ordinary ridge estimator doesn’t provide one solution to the linearity problem, but it is provides a set of solutions. These solutions depend on the K (ridge parameter) value an explicit ideal value for K cannot be found, however many random options have been proposed for this parameter [13].

3. Least Absolute Deviation (LAD)

The (LAD) method is considered one of the most common techniques for the robust regression. (LAD) estimates are not affected considerably by extreme value, comparing (OLS) estimates. However, less understood the behavior of (LAD) estimates; particularly concern the small samples, and inference process means less straightforward [2].

Inference in (LAD) estimation is an active area of research. Koenker and Bassett [10] suggested employing (LAD) estimation in the Wald, likelihood ratio (LR), and Lagrange multiplier (LM) tests. These methods can be used to test the coefficient significance in the regression model.

Dielman and Pfaffenberger [3] studied inference for regression based on (LAD) estimation in case data are independent and it is not conditioned to be normal, although (LAD) estimation has been proposed as an alternative to least squares regression, it is less used considerably and thus can be referred as a nontraditional technique.

That is (LAD) parameter estimates are the parameter values which minimize the:

$$\sum \varepsilon_i = \sum |Y_i - (a + bX_i)|$$

Where a and b are parameters which are the simple regression case and this can be generalized to multiple regression. in fact it is simpler as the absolute value of residual is more direct measure of the size of the residual than the squared residual. However the calculation of estimates (LAD) is more complicated, to estimates there exist several algorithms for calculating them [16], rather that the (LAD) method is powerful alternative to (OLS), especially when the data follows abnormal distribution and is subject to outliers.

The (OLS) regression method of analysis produces unbiased estimates parameter and has a minimum variance with independent data, identical, and naturally distributed. However, if unusual errors occur, the (OLS) could not be good performance; particularly if the errors follow the distribution tends to produce the outliers, hence, lot of research was aimed to develop the estimation approaches to be robust for such producing outlier-error distribution [2].

The (LAD) method is considered to belong to the robust regression category and can be defined as a family of regression techniques and not a single regression technique, where the robust regression is intended to be strong against outliers thus tackling potential problems, whereas, the dataset may be not always comply with the requirements of the (OLS), especially, when the high kurtosis or skew under these conditions, the regression of the (OLS) may be less efficient than some particularly robust regression method. (LAD) may be more efficient than the (OLS) method when the residual distribution is skewed or high kurtosis (LAD) is less affected by these outliers than (OLS) [14].

3.1 Least Absolute Deviation Ridge Regressions

Though (LAD) method of estimation is robust, there is still the possibility of having strong multicollinearity between the independent variables in a linear regression analysis application of (LAD) estimation instead of the (OLS) estimation for the ridge regression. The problem of multicollinearity naturally exist in most of the real life data sets, the issue is to which extent to solve the strong and middle of multicollinearity problem must be solved in one way or another.
From the linear regression model, let us take a full rank augmented linear model like
\[
\begin{bmatrix}
Y_0 \\
\end{bmatrix} = \begin{bmatrix} X \sqrt{kI_p} \beta + \epsilon \end{bmatrix}
\] (4)
with \( E(\epsilon) = 0, E(\epsilon\epsilon') = \Sigma, E(\epsilon\nu') = 0 \), where \( \nu \sim (0, \sigma^2) \), where \( k \geq 0 \) is a ridge parameter.

The (LAD) estimation of \( \beta \) in eq (4) is
\[
\hat{\beta}_r = \arg \min_{\beta=(\beta_1, \ldots, \beta_p)'} \left( \sum_{i=1}^{n} |y_i - x_i'\beta| + \sqrt{k} \sum_{j=1}^{p} |\beta_j| \right)
\] (5)
In this paper, the ridge parameters depend on the estimated parameter of the canonical form of the variables in the regression model (1)

\[
Y = X\beta + \epsilon = S'\Lambda^{1/2}U' + \epsilon
\] (6)
where the singular value decomposition of the matrix is \( X = S'\Lambda^{1/2}U' \), with \( S \) a p\times n orthogonal matrix, \( \Lambda \) a p\times p diagonal matrix of eigenvalues of \( XX' \) and \( U \) a p\times p orthogonal matrix of eigenvectors of \( XX' \), and \( \alpha = U'\beta \) the (LAD) estimation of \( \alpha \) is

\[
\begin{bmatrix}
Y_0 \\
\end{bmatrix} = \begin{bmatrix} 1_n \\
\end{bmatrix} \begin{bmatrix} S'\Lambda^{1/2} \\
\end{bmatrix} \begin{bmatrix} a_0 \\
\end{bmatrix} + \epsilon
\] (7)

In general, the optimum \( K \) differs from LAD-ridge to OLS-ridge, however, a LAD-based \( K \) and an OLS-based are aimed to find the same optimal value of \( K \) for the LAD-ridge estimation. Likewise, if they are used to estimate the OLS-ridge, then the aiming is to get a single optimal value of \( K \) for both. Therefore, the MSE for each of the bias parameters can be found in the estimation method for both the LAD-ridge and the OLS-ridge, when these biasing parameters (the LAD-based version and the OLS-based version) are taken the estimators of the optimal \( K \) value used for the OLS-ridge or the LAD-ridge estimation method.

In general, LAD-ridge estimators are more robust than OLS-ridge estimators by efficiency of LAD-ridge and least (MSE) of the biasing parameter does not always imply the least total MSE of the regression parameter estimators. The reason can be a larger intercept bias compared to the slopes bias.

Many of solutions to solve the problem of multicollinearity, and one of these solutions is ridge regression, which is considered one of the methods of bias estimating, where the ridge parameter works to give a greater area for bias in order to reduce (MSE). Here are some of ridge estimators of (LAD) estimator that were previously proposed.

3.2 Ridge Parameter
The least absolute deviation ridge estimator does not provide a single solution to the problem of linearity, but it does provide a set of solutions, these solutions depend on the value of \( K \) (ridge parameter). An explicit ideal value for \( K \) cannot be found, however, many random options have been proposed for this ridge parameter. Some of these options can be summarized as follows:-
Hoerl and Kennard (1970) \[8\] that the value of \( K_i \) which minimizes the MSE(\( \hat{\alpha}(K) \)) where:-

\[
\text{MSE}(\hat{\alpha}(K)) = \sigma^2 \sum_{i=1}^{p} \frac{1}{(\lambda_i + k)} + \sum_{i=1}^{p} \frac{k_i^2\sigma_i^2}{(\lambda_i + k)^2}
\]

Where \( K_i = \frac{\sigma_i^2}{\alpha_i^2} \), \( \sigma^2 \) represents the error variance of linear regression model, \( \alpha_i \) is the \( i^{th} \) element of \( \alpha \) and \( \alpha = U'\beta \) where \( U \) represent the orthogonal matrix and \( \lambda_i \) represent the eigenvalues.

Hocking et al (1976) \[7\] shows for the known optimal \( K_i \), the generalized ridge regression estimator is superior to all other estimators in the class of biased estimators they considered. However, the optimal value of \( K_i \) fully depends on the unknown \( \sigma^2 \) and \( \alpha_i \), and they must be estimated from the observed data. Hoerl and Kennard (1970) \[8\], suggested to replace \( \sigma^2 \) and \( \alpha_i^2 \) by their corresponding unbiased (LAD) estimators as:

\[
\hat{K_i} = \frac{\sigma_i^2}{\hat{\alpha_i}^2}
\] (8)

Where \( \hat{\sigma}^2 = \frac{\hat{\epsilon}^2}{(n-p-1)} \) is the residual mean square estimate, which is unbiased (LAD) estimator of \( \sigma^2 \) and \( \hat{\alpha_i} \) is the \( i^{th} \) element of \( \hat{\alpha} \), which is an unbiased (LAD) estimator of \( \alpha \).
B M Gomlam Kibria (2003) proposed to estimate $K$ by using the geometric mean of $\hat{K}_i$ in eq (8), as the following estimator:

$$\hat{K}_{GM} = \frac{\sigma^2}{\left(\Pi_{i=1}^{p} \hat{\alpha}_i^2\right)^{\frac{1}{p}}}$$

B M Gomlam Kibria (2003) proposed to estimate $K$ by using the median of $\hat{K}_i$ in eq (8), which produces the following estimator for $P \geq 3$:

$$\hat{K}_{MED} = \text{Median}\left\{\frac{\sigma^2}{\hat{\alpha}_i^2}\right\}, i=1,2,\ldots,p$$

Where $p$ represents the number of independent variables and its observed from eq (9) and (10) that for letting $\hat{\alpha}_i = \hat{\alpha}(i=1,2,\ldots,p)$.

### 4. Proposed ridge estimator

A method will be proposed here for ridge parameter for (LAD) estimation that includes the condition number, denoted by the symbol CN, which can be defined as the ratio between the largest to the smallest singular value, that if the condition number is very large, the matrix is said to be ill conditioned. If it is infinite then the determinant of the matrix is equal to zero [1].

The proposed estimator denoted by $\hat{K}_{CN}$ is defined as follows:

$$\hat{K}_{CN} = \text{Max}\left[0, \text{Median}\left\{\frac{\sigma^2}{\hat{\alpha}_i^2}\right\} - \frac{1}{CN}\right], i=1,2,\ldots,P$$

Where $CN$ referred to condition number, $P$ represents independent variables number.

The proposed estimator is modification $\hat{K}_{MED}$ in eq. (10) by subtracted $\frac{1}{CN}$ from $\hat{K}_{MED}$. This amount, however, varies with the strength of multicollinearity in the model. If the condition number is too large than $\hat{K}_{CN}$ would coincide with $\hat{K}_{MED}$ since in such case, the fraction $\frac{1}{CN}$ would approach to zero.

On the other hand if the condition number is too small (approximately equal to one) then the possibility that $[\text{Median}\left\{\frac{\sigma^2}{\hat{\alpha}_i^2}\right\} - \frac{1}{CN}, i=1,2,\ldots,p]$ be negative is too large.

If it is equal to zero which means that the ridge regression estimator would coincide with the least absolute deviation and the data set is not influenced by the multicollinearity problem [13].

### 5. Simulation Study

In this section we will discuss a simulation study to compare the performance of the ridge estimators in order to find out which of the ridge parameters are most efficient and having least (MSE). In this research, three independent variables $P=3$ were used in the linear regression model that represents the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i=1,2,\ldots,n$$

McDonald and Galarneau (1975) [12] and Gibbons (1981) [4] the independent variables were generated by using the following device:

$$x_{ij} = (1 - \rho^2)^{\frac{i}{2}} \epsilon_{ij} + \rho \epsilon_{ij}, \quad i=1,2,\ldots,n, \quad j=1,2,\ldots,p$$

Where $\epsilon_{ij} \sim N(0,\sigma^2)$ is independent normal and $\rho$ represents the correlation between two independent variables which is inactivated by $\rho^2$ in this paper it is assumed that it is $\beta_0 = \beta_1 = \beta_2 = \beta_3 = 1$, then the following correlation values were (0.3, 0.6, 0.9) different sample sizes were used $n=25,50,200$ and the value of $\sigma^2$ is (0.5, 1, 2). The MATLAB (2020 a) program was used in order to compare the two ridge parameters to show which one is the best. Tables below represents the MSE values the best parameter with smallest MSE.

| $\sigma^2$ | $\hat{K}_{MED}$ | $\hat{K}_{CN}$ | $\hat{K}_{MED}$ |
|------------|-----------------|----------------|----------------|
| 0.5        |                 |                |                |
| 1          |                 |                |                |
| 2          |                 |                |                |

Table 1. The values of MSE at $\rho=0.3$.
Table 2. The values of MSE at ρ=0.6.

| Estimator | σ²  | MSE |
|-----------|-----|-----|
| $\hat{K}_{MED}$ | 0.022 | 0.004 |
| $\hat{K}_{CN}$ | 0.032 | 0.009 |
| $\bar{K}_{MED}$ | 0.044 | 0.018 |
| $\bar{K}_{CN}$ | 0.054 | 0.021 |

Table 3. The values of MSE at ρ=0.9.

| Estimator | σ²  | MSE |
|-----------|-----|-----|
| $\hat{K}_{MED}$ | 0.042 | 0.006 |
| $\hat{K}_{CN}$ | 0.052 | 0.010 |
| $\bar{K}_{MED}$ | 0.064 | 0.018 |
| $\bar{K}_{CN}$ | 0.074 | 0.021 |

We also worked on showing the results by figures. Where the figures (1,2,3,4,5,6) below represent the MSE values for each of the two ridge parameters ($\hat{K}_{MED}$ and $\hat{K}_{CN}$).

**Figure 1.** The behavior of $\hat{K}_{MED}$ when ρ=0.3.

**Figure 2.** The behavior of $\hat{K}_{CN}$ when ρ=0.3.
5.1 Analysis of simulation results

From table (1,2 and 3) which contains the simulation results of MSE of ridge parameter (K), we have observed that:

When $\sigma^2=0.5$ in experiment 1 and $\rho=0.3$, the best ridge parameter is $\hat{K}_{MED}$ in both sample size (25 and 50), both in large sample size (n=200) the best is $\hat{K}_{CN}$, where in experiment 2 when $\rho=0.6$ the best is $\hat{K}_{CN}$.
When $\sigma^2=1.2$ and in the two experiments in $\rho=0.3,0.6$, we observe that almost in smaller sample size the best is $\hat{K}_{MED}$ and in larger sample size the best ridge parameter is $\hat{K}_{CN}$.

In experiment 3 and when $\rho=0.9$ the best is $\hat{K}_{CN}$ in all values of $\sigma^2$ and all sample sizes, this results confirms the effectiveness of the proposed estimator $\hat{K}_{CN}$.

Through the figures (1,2,3,4,5,6), we have observed that:
- $\hat{K}_{MED}$ at $\rho=0.3$ has the smallest (MSE) when $\sigma^2=1$.
- $\hat{K}_{MED}$ at $\rho=0.6$ has the smallest (MSE) when $\sigma^2=2$.
- $\hat{K}_{CN}$ at $\rho=0.3$ have smallest (MSE) when $\sigma^2=0.5$.
- $\hat{K}_{CN}$ at $\rho=0.6$ has the smallest (MSE) when $\sigma^2=0.5$.

We also observed that $\hat{K}_{MED}$ and $\hat{K}_{CN}$ behavior have smallest (MSE) when $\sigma^2=2$.

6. References

[1] Cassela, G. (1985); "Condition Numbers and Minimax Ridge Regression Estimators". *Journal of The American Statistical Association*, Vol. 80, No, 391, PP753-758.

[2] Dadamos, T. R. L, Damaceno, A. J. Fertonani, F. L., and Bettencourt Silva, R. J. N. (2018) "Use of Monte Carlo Simulations for The Calculation of The Uncertainty Estimate in The Electrochemical Detection of Uric Acid in Plasma Samples". *Periodico Tchê Quimica*, 15(29), PP. 196-203.

[3] Dielman, T. E. and Pfaffenberger, R. (2009) "Tests of Linear Hypothesis in LAV Regression". *Communications in Statistics-Simulation and Computation*, 19, PP. 1179-1199.

[4] Gibbons, D. G. (1981). "A Simulation Study of Some Ridge estimators". *Journal of The American Statistical Association* 76:131-139.

[5] Hoerl, A. E. (1962). "Application of Ridge Analysis to Regression Problems". *Chemical Engineering Progress* 58:54-59.

[6] Hoerl, A. E., Kennard, R. W., Baldwin, K. F. (1975). "Ridge Regression: Some Simulation". *Communications in Statistics* 4:105-123.

[7] Hocking, R. R., Speed, F. M., Lynn, M. J. (1976). "A Class of Biased Estimators in Linear Regression". *Technometrics* 18:425-438.

[8] Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression Biased Estimation for Non Orthogonal Problems", *Techno Metrics, Vol. 12, PP 55-67.

[9] Kibria B. M. G., (2003). "Performance of Some New Ridge Regression Estimators", *Communication in Statistics-Theory and Methods*, 32:419-435.

[10] Koenker, R. and Bassett, G. (2002). "Tests of Linear Hypothesis and L1 estimation". *Econometrical, 50, PP. 1577-1583.

[11] Lawless, J. F., Wang, P. (1976). "A Simulation Study of Ridge and Other Regression Estimators". *Communications in Statistics A5:307-323.

[12] McDonald, G. C., Galarneau, D. I. (1975). "A Monte Carlo Evaluation of Some Ridge-Type Estimators". *Journal of the American Statistical Association* 70:407-416.

[13] Mahdi, Fatima Assim (2017). "A Comparison Different Ridge Regression Methods with Application". University of Baghdad-College of Education for Pure Science/ Ibn Al-Haitham/ Baghdad/ Iraq.

[14] Pickard, L., Kitchenam, B. and Linkman, S. "An Investigation of Analysis Techniques for Software Data Sets". *Proc. Metrics 99 Symposium, IEE Computer Society, 1999, PP. 130-142.

[15] Pfaffenberger, R.C., and Dielman, T. E., (1989). "A Comparison of Regression Estimators When Both Multicollinearity and Outliers and Present". In: Lawrence, K. D. and Arthur, J. L., Editors, Robust Regression: Analysis and Applications. New York: Marcel Dekker, Inc, PP. 243-270.

[16] Tronfass, Ingunn Myrteit, Erikstensrud, (2001). "A Comparison of LAD and OLS Regression for Effort Prediction of Software Projects".

[17] Zangin Zeebari (2012). "A Simulation Study on the Least Absolute Deviations Method for Ridge Regression". *Department of Economics and Statistics, Jönköping International Business School, Sweden, Article in Communication in statistics-theory and method."