Global Scientific JOURNALS

GSJ: Volume 9, Issue 12, December 2021, Online: ISSN 2320-9186

www.globalscientificjournal.com

ROBUST RIDGE ESTIMATORS FOR HANDLING MULTICOLLINEARITY AND OUTLIERS IN LINEAR REGRESSION MODELS

[1]Adamu Tijjani*, [2]Yusuf A. Mohammed, [3]Adeniyi Ogunmola Oyewole and [4]Nicholas P. Dibal

[1,2,4]Department of Mathematical Science, University of Maiduguri, P.M.B 1069 Maiduguri, Nigeria.

[3]Federal University Wukari, Taraba State, Nigeria

.

**Corresponding Author:tujjaniadamu@gmail.com,*

yusufabbakarm@gmail.com, adeniyiogunmola@mail.com, pndibal@unimaid.edu.ng

Abstract

In multiple linear regression analysis, the ordinary least squares (OLS) method has been the most popular technique for estimating parameters of linear regression model due to its optimal properties. OLS estimator may fail when the assumption of independence is violated. This assumption can be violated when there is correlation between the exploratory variables. In this situation, the data is said to contain multicollinearity and eventually will mislead the inferential statistics. However, the problem becomes more complicated when there are abnormal observational data known as outliers. Multicollinearity and Outliers are two main problems. When multicollinearity exists, biased estimation technique Ridge Estimator are preferable to OLS. On the other hand, when outliers exist in the data, robust estimator like LTS Estimator, are preferred. To handle these two problems jointly, the study ussed Robust Ridge Regression estimator. This study aims to examine the performances of four estimators of multiple linear regression model with combined problems of multicollinearity and outliers. The performance of the four estimators, namely the Ordinary Least Squares (OLS), Ridge Regression (RIDGE), Least Trimmed Square (LTS) and a robust ridge regression estimator based on Least Trimmed Square estimator (RIDGE LTS) are compared using mean square error as criteria for assessment. For this purpose, a simulation data with p = 3; n = 25, 50,100; and full multicollinearity (r = 0.90, 0.95, 0.99) and outliers (0%, 20%) was used. The existence of multicollinearity was evaluated using VIF value. The empirical evidence shows that RIDGE LTS is the best among the four estimators for degree of multicollinearity and number of outliers and is more efficient because it has the smallest MSE than LTS and RIDGE in any samples sizes.

Keywords

OLS, Ridge Regression, LTS Estimator, Ridge Estimator, Multicollinearity, Outliers, MSE.

1.0 Introduction

Multiple linear regression is a widely used statistical techniques in the study of relationship between a single response variable Y, and one or more explanatory variable(s) $X_1, X_2, \ldots, X_p$ using linear models. Ordinary Least Squares (OLS) is a method commonly used in estimating the parameters of linear regression models, it is said to be the Best Linear Unbiased Estimator (BLUE) when all the required assumptions are satisfied. However, in real life situations these assumptions are hardly satisfied thereby

influencing parameter estimates negatively. The major problems faced in multiple linear regression analysis are the issues of multicollinearity and outliers are the two main problems. When multicollinearity exists, the OLS estimation is seriously affected, as such, unbiased estimation techniques such as Ridge (Hoerl and Kennard, 1970) and Liu (Liu, 1993) Estimators are preferable to Ordinary Least Square. On the other hand, when outliers exist in the data, robust estimators like M, MM, LTS and S Estimators, are preferred (Adegoke *et al*., 2016). Multiple linear regression model is commonly used to determine the best set of parameters of the linear model so that the predicted value of dependent variables approaches the real values (Fitrianto, A. and Yik, L.C., 2014). The assumption of independence of the predictor variables are normally violated in real life situation and predictor variables are found to be correlated. This inter-relation between the explanatory variables is called multicollinearity. Multicollinearity, or collinearity, is the existence of near linear relationships among the independent variables, its presence creates inaccurate estimates of the regression coefficients, inflate the standard errors of the regression coefficients, deflate the partial t-tests for the regression coefficients, give false, nonsignificant and degrade the predictability of the model. The source of multicollinearity must be identified to be able to reduce its effect on the analysis and interpretation of the linear model. The ridge regression is a regression technique that allows for biased estimation of regression parameters that are quite close to the true values in the presence of correlated predictor variables in the model. All the various forms of the ridge regression techniques were meant to shrink the least square coefficients towards the origin of the parameter space and consequently reduce the mean square errors of estimates (Yahya and Olaif, 2014). Roozbeh *et al*. (2021) proposed a robust ridge test statistic which was used to improve the predictions in a regression model, they introduced robust ridge type estimator in the presence of multicollinearity and outliers which improves it by shrinking toward the origin to incorporate the information contained in the null-hypothesis. Zhang and Mahmud (2021) compare the ordinary least squares (OLS) regression and ridge regression procedures in dealing with multicollinearity data since it is well known that the LS method is extremely unreliable in parameter estimation while the independent variables are dependent (multicollinearity problem).

When multicollinearity and outliers exist together in a data set, robust approach is suggested. The methods of ridge regression and robust estimator are combined to handle the problems jointly (Montgomery and Peck, 1982). This study investigates the efficiency of some estimators; these are ridge least trimmed square (RLTS), least trimmed square (LTS), and ridge regression (RR) in the presence of multicollinearity and outliers against ordinary least squares estimator (OLS) using Mean square error (MSE) as model's assessment criteria for examining the performance of the estimators. This paper is organized as follows: methodology in Section 2 followed by simulation in Section 3, data analysis in Section 4 and the concluding remarks in the last section.

2.0 Methodology

The ordinary least squares (OLS) for estimating the parameters of a linear model, the least trimmed squares (LTS) which is a robust estimator of parameters in contaminated data and the ridge regression estimator were used followed by robust ridge estimator.

i) Ordinary Least Squares (OLS)

Ordinary least square is a statistical technique used to estimate the parameters of a linear regression equation. The aim of this technique is to determine the line of best fit for the given data by minimizing the sum of squared errors. The standard regression model is represented by matrix notation of a multiple linear regression model:

$$Y = X\beta + \varepsilon \quad (1)$$

Note that both X's and Y have been standardized. The OLS estimate of $\widehat{\beta}$ of β is obtained by minimizing the residual sum of squares as:

$$\hat{\beta}_{OLS} = (X^T X)^{-1} X^T X \quad (2)$$

$$var(\hat{\beta}_{OLS}) = \hat{\delta}^2 (X^T X)^{-1} \quad (3)$$

where $\hat{\sigma}^2$ is the mean squares error. This estimator $\hat{\beta}$ is unbiased and has a minimum variance.

ii) Ridge Regression Method

The least-squares method gives weak estimates of the regression coefficients on non-orthogonal data (Hoerl and Kennard, 1970). Ridge regression estimator is a method that is more efficient than ordinary least squares when the data exhibit multicollinearity. The method works by adding a scalar ridge parameter which is called the biasing constant to the main diagonal of matrix. In the presence of multicollinearity, the is singular and may result in poor estimates. Here, the parameter is then added to improve the matrix condition before computing $_s$ by using method of Hoerl and Kennard (1970). The ridge solution is given by:

$$\widehat{\beta}_{RIDGE} = (X^I X + kI)^{-1} X^I Y \qquad k \geq 0 \quad (4)$$

Where I is (p × p) identity matrix and is a scalar ridge parameter. Various methods have been introduced to obtain the value of k which is the main challenge in ridge regression. In general, Ridge Trace, Generalized Cross Validation methods is commonly used, however the Hoerl, Kennard and Baldwin (1975) method were employed in this study to find.

$$k_{HKB} = \frac{PS_{LS}^2}{\beta_{LS}^I \widehat{\beta}_{LS}} \quad (5)$$

where $s_{LS}^2 = \frac{(Y - X\widehat{\beta}_{LS})^I (Y - X\widehat{\beta}_{LS})}{n-p}$ (6)

This method reduces MSE of the regression parameter by adding a positive value of ridge parameter, such that an increase in the bias lead to the reduction of the variance. Note that if k = 0, the ridge estimator (4) is name as the OLS (2) which means $\hat{\beta}_R = \hat{\beta}_{OLS}$. When the value of k > 0, MSE($\hat{\beta}_R$) < MSE($\hat{\beta}_{oLS}$), $\hat{\beta}_R$ is biased but will be more precise and stable for the ridge regression estimator (4). It is also true that $\widehat{\beta}_R \rightarrow 0$ when the estimator k → ∞. If all $\hat{k_s}$ are the same the resulting estimators are called the ordinary ridge estimators.

iii) Least Trimmed Square (LTS)

The most used robust estimator is the Least Trimmed Square (LTS) proposed by (Rousseuw, 1984) which is used to fit a regression by using estimators that dampen the effect of influential points in both dependent and the explanatory variables, however the LTS estimates perform badly on a normally distributed data. The estimator is very robust to the presence of outliers. This estimator minimizes the sum of trimmed squared residuals and is written as;

$$\hat{\beta}_{LTS} = \min\sum_{i=1}^{h} \varepsilon_i^2(\beta) \quad (7)$$

Such that h= $\frac{n}{2}$ + $\left(\frac{p+1}{2}\right)$ with n and p being sample size and number of variables respectively, and, $\varepsilon_{(1)}^2 \leq \varepsilon_{(2)}^2 \leq \varepsilon_{(3)}^2 \leq, \ldots, \leq \varepsilon_{(n)}^2$, the ordered squared residuals. LTS estimator may be very efficient based on the value of and the outliers. The largest squared residuals are being excluded from the summation in this method. Therefore, it allows those outlier data points to be excluded completely. Contradictory, LTS estimator may not be efficient if the number of trimmed data points is more than actual outliers as some good data will be excluded. Furthermore, if the exact numbers of outliers in the data set are trimmed, this method of calculation is like the OLS.

iv) Robust Ridge Regression Estimator

When both outlier and multicollinearity occur in a data set, it would seem preferred to combine methods for dealing with these problems simultaneously. To illustrate the idea of how the combination of ridge regression and robust estimators work, for robust ridge regression by combining the properties of the least trimmed square (LTS) and the ridge regression estimator referred to as Ridge trimmed square (Ridge LTS) estimator:

$$\hat{\beta}_{RLTS} = (X^T X + k_{LTS} I)^{-1} X^T Y \quad (8)$$

where the value of k is determined using:

$$k_{LTS} = \frac{p s_{LTS}^2}{\beta_{LTS}^I \hat{\beta}_{LTS}} \quad (9)$$

And $$s_{LTS}^2 = \frac{(Y - X\hat{\beta}_{LTS})^I (Y - X\hat{\beta}_{LTS})}{n-p} \quad (10)$$

where p is the number of parameters, n is the sample size and s_{LTS}^2 is the estimated variance.

3.0 Simulation Study

Simulation study was conducted to assess the performance of Ordinary Least Square (OLS), Ridge Regression (Ridge), Least Trimmed Square (LTS) and Ridge Least Trimmed Square (Ridge LTS) estimators. The simulation study was designed with three levels of high correlation (r=0.90, 0.95, 0.99) between the explanatory variables and three sample sizes (n=25, 50, 100) was used. The standard normal distribution is used with 1000 trials (replication) for each sample size. The percentages of outlier injected is (0%, 20%). The simulation model used for the study is:

$$y_i = \beta_o + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i \quad (11)$$

The explanatory variables $x_{i1}, x_{i2\ and}\ x_{i3}$ were generated using the following equation;

$$x_{ij} = (1 - r^2)z_{ij} + rz_{ij} \quad for \quad i = 1,2,\dots,n \quad j = 1,2,3\ and \quad (\beta_j = 1\ for\ j = 0,1,2,3) \tag{12}$$

where z_{ij} are independent standard normal random numbers that is held fixed for given sample of size n, and r is the theoretical correlation between the x_{ij} and are fixed at 0.90, 0.95 and 0.99. The statistical computation of MSE is given by

$$MSE = \frac{\sum_i^n (y_i - \hat{Y}_i)^2}{n} \tag{13}$$

where, y_i are the true values, $\hat{Y}_i$ is the predicted values and n is the sample size,

Performance Measures of the Estimators

Using a simulated data set with multicollinearity and outliers, the proposed estimators were compared using the Mean Square Error (MSE) criterion.

4.0 Results

Full multicollinearity is designed into independent variables. The condition is analyzed using VIF values of the variables, which are used to diagnose multicollinearity problems. The VIF can help determine which regressors are implicated in the multicollinearity problem.

Table 4.1 The VIF values of Independent Variables

		Variance Inflation Factor (VIF)		
Variable	Sample Size (n)	r = 0.90	r=0.95	r=0.99
×1	25	3.3431	17.9207	97.4527
	50	4.4861	8.9639	27.4952
	100	4.0101	45.3533	39.7301
x2	25	4.9201	27.0229	1340.7848
	50	10.6016	21.1438	55.0810
	100	4.0142	39.6109	882.2348
x3	25	4.1095	17.3606	1870.7670
	50	7.4700	14.3951	14.3951
	100	1.0138	10.0535	113.7319

Except where the correlation, the VIF of the variables in Table 4.1 is more than 5. The maximum VIF is 1870.7670, indicating that the correlation between independent variables is extremely high, indicating that all three independent variables are fully multicollinear. As a result, the multicollinearity problem is undeniable.

Table 4.2. VIF after the Application of Ridge Regression (k=1)

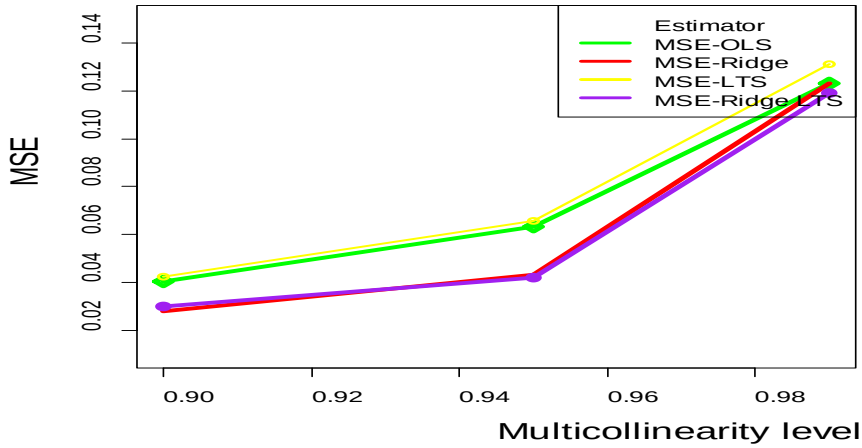
		Variance Inflation Factor (VIF)		
Variable	Sample Size (n)	r = 0.90	r= 0.95	r= 0.99
×1	25	0.1623	0.0880	0.0707
	50	0.1464	0.1103	0.0960
	100	0.1655	0.0781	0.0804
x2	25	0.1384	0.0793	0.0661
	50	0.1053	0.0853	0.0737
	100	0.1654	0.0821	0.0693
x3	25	0.1507	0.0805	0.0641
	50	0.1264	0.0993	0.1067
	100	0.2492	0.1029	0.0664

The VIF is reexamined after the ridge regression is performed to the data to see if the multicollinearity problem has been overcome. The results reveal that when ridge regression is used, the VIF values drop dramatically approaching one. It shows that ridge regression is quite good at dealing with multicollinearity.

Table 4.3 Estimated MSE For OLS, Ridge, LTS, Ridge.LTS Estimators for Different Sample Sizes and levels of Multicollinearity Without Outliers

		Mean Square Error (MSE)		
Sample Size (n)	Estimator	r=0.90	r=0.95	r=0.99
25	OLS	0.0405	0.0632	0.1233
	Ridge	0.0280	0.0430	0.1230
	LTS	0.0423	0.0658	0.1310
	Ridge.LTS	0.0300	0.0420	0.1190
50	OLS	0.0144	0.0170	0.0244
	Ridge	0.0070	0.0090	0.0190
	LTS	0.0125	0.0148	0.0213
	Ridge.LTS	0.0060	0.0080	0.0160
100	OLS	0.0073	0.0122	0.0160
	Ridge	0.0070	0.0080	0.0160
	LTS	0.0079	0.0157	0.0172
	Ridge.LTS	0.0070	0.0070	0.0150

When data is simulated using sample sizes of 25, 50, and 100, Table 4.3 shows the relative performance of the estimators in the presence of three different levels of multicollinearity.



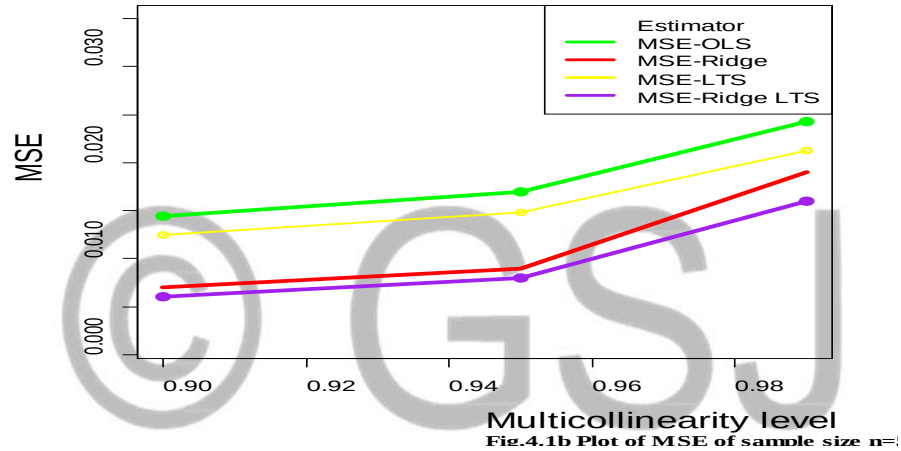
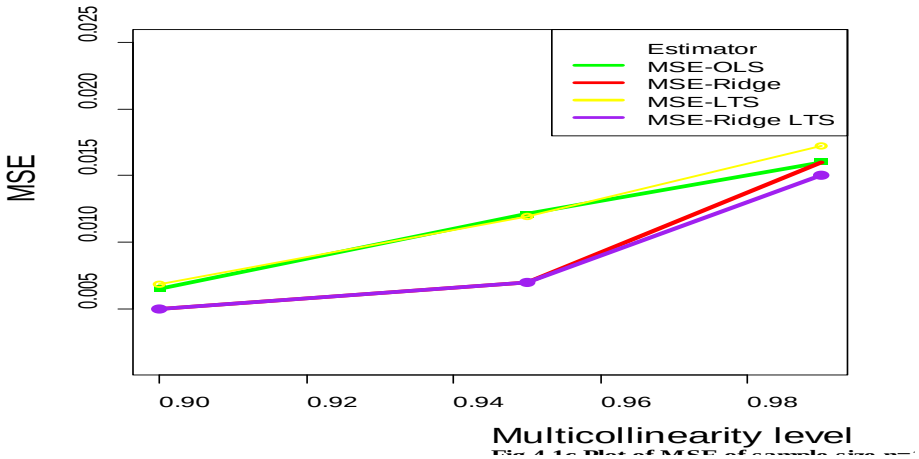


Fig.4.1b Plot of MSE of sample size n=



The MSE of the Ridge and Ridge LTS are shown in Table 4.1. When the errors are normally distributed and multicollinearity is present at a correlation value of r = 0.90, Ridge LTS is smaller than the other estimators. The result in Table 4.2 favors Ridge and Ridge LTS at r = 0.95. LTS is used for normal error distributions when collinearity is present in the data. Ridge is better than OLS, LTS, and its performance is

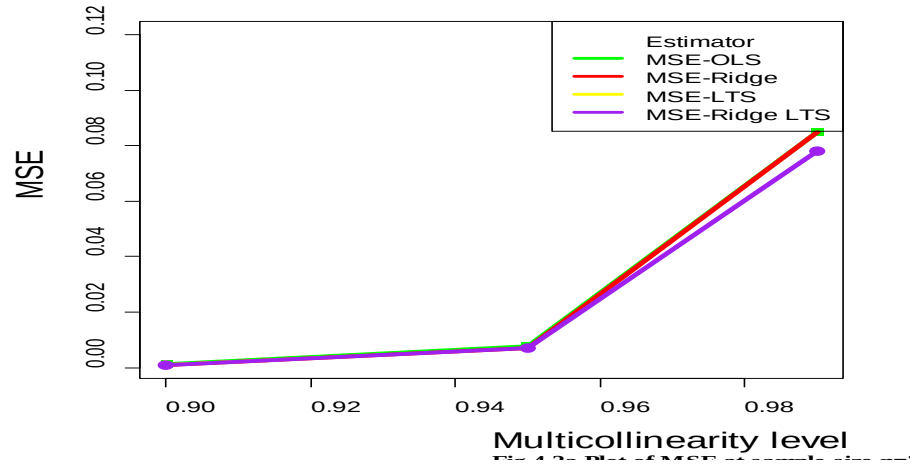
virtually as excellent as Ridge, according to the MSE in Table 4.2, which corroborated the conclusion from Table 4.3. At r = 0.99, LTS has a high collinearity level. Ridge LTS is superior in every other way. The simulation results for bigger samples, n=100, are, nonetheless, consistent with the results for smaller samples. Because the MSE values are smaller, the results suggest that estimates for bigger samples are more efficient than those for smaller samples. For further clarity, the MSE values were recorded in tables 4.1, 4.2, and 4.3, and the data were presented in Figs. 4.1a, 4.1b, and 4.1c, respectively. Ridge LTS estimators were found to be the best estimators since they have the lowest values of the criterion considered in the assessment. Furthermore, as the level of multicollinearity was raised, the estimators' performance deteriorated.

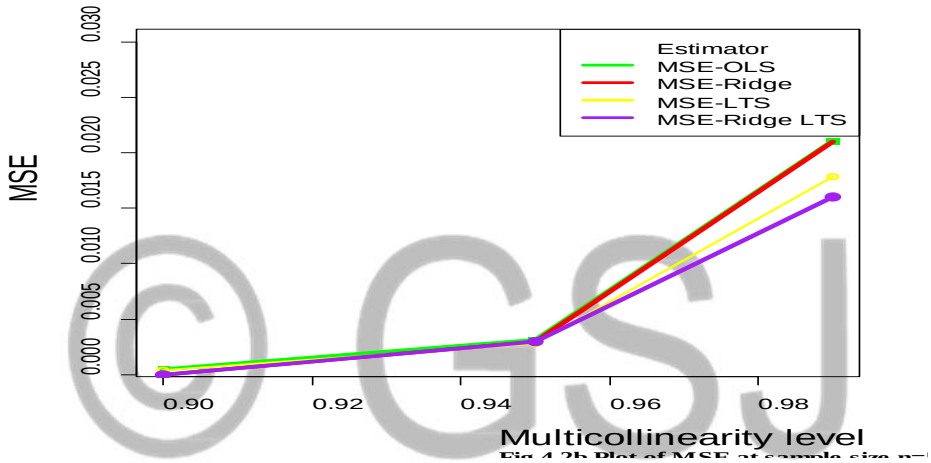
4.4 Effect of Multicollinearity and Outliers on the Estimators.

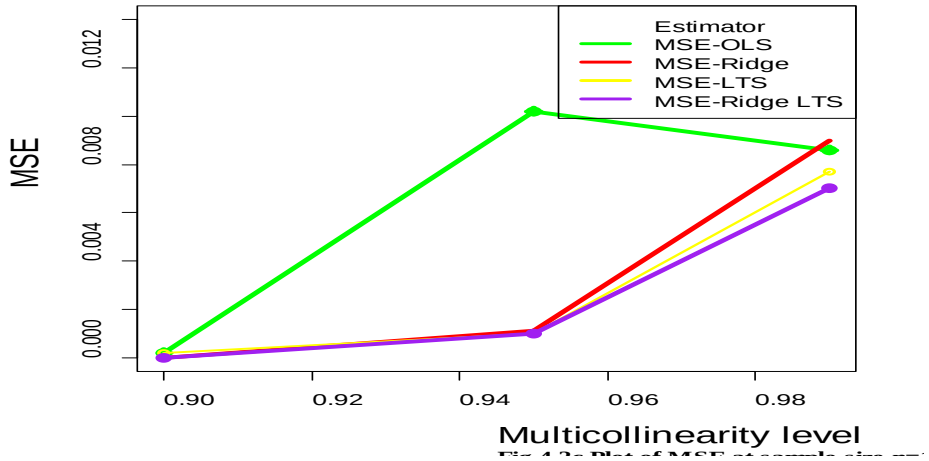
When data are simulated using sample sizes of 25, 50, and 100 with 20% outliers, Table 4.4 shows the relative performance of the estimators in the presence of three distinct levels of multicollinearity.

Table 4.4 Estimated MSE For OLS, Ridge, LTS, Ridge.LTS Estimators for Different Sample Sizes, levels of Multicollinearity and 20% Outliers

		Mean Square Error (MSE)		
Sample Size (n)	Estimator	r=0.90	r=0.95	r=0.99
25	OLS	0.0012	0.0078	0.0849
	Ridge	0.0010	0.0070	0.0850
	LTS	0.0011	0.0071	0.0779
	Ridge.LTS	0.0010	0.0070	0.0780
50	OLS	0.0004	0.0031	0.0211
	Ridge	0.0000	0.0030	0.0210
	LTS	0.0004	0.0028	0.0179
	Ridge.LTS	0.0000	0.0030	0.0170
100	OLS	0.0001	0.0011	0.0086
	Ridge	0.0000	0.0010	0.0090
	LTS	0.0002	0.0010	0.0077
	Ridge.LTS	0.0000	0.0010	0.0070







For data with a non-normal error distribution and multicollinearity, and for each sample size and number of outliers, Tables 4.4, 4.5, and 4.6 indicate that Ridge LTS produces the least MSE value, followed by LTS and OLS. In any number of sample sizes, Ridge LTS handles multicollinearity and the number of outliers substantially better than Ridge, OLS, and LTS. For further clarity, the MSE values were recorded in table

4.4, 4.5, and 4.6, and the data were presented in Fig. 4.2a, 4.2b, and 4.2c, respectively. Ridge LTS estimators were found to be the best estimators since they have the lowest values of the criterion considered in the assessment.

Furthermore, as the level of multicollinearity was raised, the estimators' performance deteriorated. The results of simulations for bigger samples are like those of smaller samples. The results also show that the estimator for bigger samples is more efficient than for smaller samples, as evidenced by the lower MSE values. As a result, when the errors are uniformly distributed in the presence of multicollinearity and outliers, the MSE of the Ridge LTS is smaller than the other estimators. When multicollinearity and outliers are present, Ridge LTS is more efficient than LTS and Ridge, and certainly much more efficient than OLS.

5.0 Conclusion

The MSE derived via Ridge LTS is the lowest, as can be shown. When both multicollinearity and outliers are present, simulation tests clearly reveal that the Ridge LTS estimate is the most practical option over other estimators. The MSE value for a large sample size is much lower than for a small sample size, implying that a larger sample size produces better and more reliable results. As the sample size grows, the case results of the estimation methods become more stable. It can be concluded that Ridge LTS is a better method for handling multicollinearity and outliers than OLS, Ridge, and LTS for small and large sample sizes.

References

Adegoke, A. S., Adewuyi, E., Ayinde, K. and Lukman, A. F. (2016). A Comparative Study of Some Robust Ridge and Liu Estimators. *Science World Journal.* **11** (4). 16-20 *www.scienceworldjournal.org*

Fitrianto, A. and Yik, L.C. (2014). Performance of ridge regression estimator methods on small sample size by varying correlation coefficients: a simulation study. *Journal of Mathematics and Statistics*. 10 (1): 25-29, doi:10.3844/jmssp.2014.25.29

Hoerl, A.E and Kennard, R.W. (1970) Ridge regression: Biased estimation for non-orthogonal problems, Technometrics, 12(1). 55-67

Hoerl, A. E, Kennard, R. W., and Baldwin, K. F. (1975). Ridge regression: Some simulations Communications in Statistical Theory, **4** (2), 105-123.

Liu, K. (1993). A new class of biased estimate in linear regression. *Communication in Statistics.* **22** (2): 393-402.

Montgomery, D.C and Peck, E.A Introduction to Linear Regression Analysis, New York; Wiley and Sons,Inc.,(1982).

Roozbeh, A. M., M, Hamzah, N. A. and Gasparini, M (2021) Ridge regression and its applications in genetic studies.PLoS ONE 16(4):e0245376. doi.org/10.1371/journal.pone.0245376

Rousseeuw, P.J. and K. Van Driessen, (1998). “Computing LTS regression for large data sets”,Technical Report, University of Antwer.

Yahya, W. B. and Olaif, J. B. (2014). A note on ridge regression modeling techniques. *Electronic Journal of Applied Statistical Analysis,* **07** (02), 343-361 DOI: 10.1285/i20705948v7n2p343

Zhang, J. and Mahmud Ibrahim (2021). A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. 571-588 doi.org//10.1080/02664760500078946.