PHYSICAL SCIENCES

ISSN (Print): 24490954
ISSN (Online): 26364972

RELATIVE EFFICIENCY OF RIDGE REGRESSION AND ORDINARY LEAST SQUARE ESTIMATORS ON LINEAR REGRESSION MODELS AT DIFFERENT LEVELS OF MULTICOLLINEARITY

Akeyede, I., *Ailobhio, D.T and Ayoo, P.V.

Department of Mathematics
Federal University Lafia, PMB 146, Lafia, Nigeria

**Corresponding Email: imamakeyede@gmail.com*

Manuscript Received: 15/03/2017 *Accepted: 23/11/2017* *Date Published: December 2017*

ABSTRACT

In Linear Regression models, multicollinearity has been observed to influence estimation of the model parameters. This study therefore examined the effect of multicollinearity on two different methods of parameter estimation, namely; Ridge Regression (RR) and Ordinary Least Squares (OLS) Estimators. A simulation technique was conducted to examine the relative efficiency of the estimators when the assumption of no multicollinearity (no correlation) between the explanatory variables is violated. Finite properties of estimators' criteria namely, absolute bias and mean squared error were used for comparing the methods. An estimator is best at a specified level of multicollinearity and sample size if it has minimum total criteria. The performance ofbothestimators for estimating the parameters of the regression models were the same when there is low level of multicollinearity in the model. However, the Ridge Regression estimator outperformed others as level of multicollinearity was increased and it is therefore recommended for the analysis of the linear regression models when there is high level of multicollinearity.

Keywords: *Ridge Regression, Ordinary Least Squares, Multicollinearity, Absolute Bias, Mean Squared Error.*

INTRODUCTION

Least squares multiple regression with a single dependent variable has been successfully applied to a variety of scientific fields. This can be attributed to the Gauss-Markov theorem, which states that the least squares estimator is the best linear unbiased estimator (BLUE). The best estimator is based on assumptions $E(\varepsilon_i) = 0,\ V(\varepsilon_i) = \sigma^2,$ and $Cov(\varepsilon_i, \varepsilon_j) = 0.$ Even if in the cases when the assumptions are not fulfilled, the least squares estimator is"best" among the unbiased and linear estimators. It is well known that thereexist estimators which are either biased or non-linear or both, which have anaverage performance far better than the least squares estimator. In regression, the objective is to explain the variation in one or more response variables, by associating this variation with proportional variation in one or more explanatory variables. One frequent obstacle is that several of the explanatory variables will vary in rather similar ways. As a result, their collective power of explanation is considerably less than the sum of their individual powers. Predictor variables $x_1, x_2, x_3, \cdots\cdots, x_p$ are assumed to be linearly independent of each other, if this assumption is violated; the problem is referred to as collinearity problem. If two or more independent variables are dependent on one another, it is called Multicollinearity(Adnan, *et al*., 2006).

This phenomenon called multicollinearity is a common problem in regression analysis. Handling multicollinearity in regression analysis is important because least squares estimations assume that predictor variables are not correlated with each other. There are various procedure for dealing with multicollinearity some of these include Indirect Least Square (ILS), Partial Least Square Regression (PLSR), Ridge Regression (RR), Lasso etc (Adnan, *et al*., 2006). The least squares estimator also fails when the number of explanatory variablesis relatively large in comparison to the sample size.

The method of least squares has some very attractive statistical properties that have made it one of the most powerful and popular methods of regression analysis.Least square estimators are unbiased and have variance co-variance matrix given by $V(\hat{\beta}) = (X'X)^{-1}\sigma^2$. Hence, $(\hat{\beta})$ is an unbiased estimate of β.

Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors. RR was developed by Hoerl and Kennard (1970) and this method is the modification of the least squares method that allows biased estimators of the regression coefficients. Although the estimators are biased, the biases are small enough for these estimators to be substantially more precise than unbiased estimators. Therefore, these biased estimators are preferred over unbiased ones since they will have a larger probability of being close to the true parameter values (Neter, *et. al*., 1985, 1996).

MATERIALS AND METHODS

Ordinary Least Squares

Following the usual notation, suppose our regression equation is written in matrix form as

$$Y = X\beta + \varepsilon \quad \text{.....(1)}$$

Where Y is the dependent variable, X represents the independent variables, β is the regression coefficients to be estimated, and represents the errors and residuals.

X $= x_1, x_2, x_3, \cdots\cdots, x_p$ where Y is a real-valued function. To predict Y form X by f(x) so that the expected loss function $E(L(Y, f(x)))$ is minimized, $f(X) = E(Y/X)$

And the function $E(Y/X)$ is the regression function. The linear regression model,

$$f(X) = \beta_o + \sum_{j=1}^{p} X_j \beta_j$$

$$\hat{Y} = \beta_o + \sum_{j=1}^{p} X_j \beta_j \quad \text{.....(2)}$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Fitted vector is $\hat{Y} = X\hat{\beta} = X(X^T X)^{-1} X^T Y$

Hat Matrix (H) is $H = (X^T X)^{-1} X^T Y$

If linear model is true, $E(Y/X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$

and the least square estimate of β is unbiased

$E(\hat{\beta}_j) = \beta_j \quad j = 0,1,\cdots, p$

$Y = E(Y/X) + \varepsilon$

Where $\varepsilon \sim N(0, \sigma^2)$ is independent of X, X_{ij} are regarded as fixed. Y_i are random due to ε.

$Var(\hat{\beta}) = (X^T X)^{-1}\sigma^2$

$\hat{\beta} \sim N(\beta, (X^T X)^{-1}\sigma^2)$

Shortcomings of OLS:

(i). OLS regression has zero bias but suffers from high variance when collinearity is presence in the regression coefficients, so it may be worth sacrificing some bias to achieve a lower variance

(ii). Interpretation become a problem when there is large number of predictors, hence it can be helpful to identify a smaller subset of important variables.

(iii). OLS regression is not defined when P>n.

Ridge Regression

Ridge regression panelizes the size of the regression coefficients. Specifically, the ridge regression estimates $\hat{\beta}$ is defined as the value of β (Khalaf and Shukur, 2005) that minimizes.

$$\sum (Y_i - X_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad \text{.....(3)}$$

The solution to ridge regression problem is given by:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y \quad \text{.....(4)}$$

Note the similarity with the ordinary least squares solution, but with the addition of a ridge down the diagonal.

As $\lambda \to 0$, $\hat{\beta}^{ridge} \to \hat{\beta}^{OLS}$

As $\lambda \to \infty$, $\hat{\beta}^{ridge} \to 0$

In special cases of an orthonormal design matrix

$$\hat{\beta}^{ridge} \to \frac{\hat{\beta}^{OLS}}{1+\lambda}$$

This illustrates the essential feature of ridge regression shrinkage. Applying the ridge regression penalty has the effect of shrinking the estimates toward zero, introducing bias but reducing the variance of the estimate (Hoerl, *et al.*,1976). And its benefit is most striking in the presence of multicollinearity.

Recall in OLS, estimates do not always exist, if X is not a full rank $(X^T X)$ is not invertible and there is no unique solution for β^{OLS} . This problem does not occur with ridge regression. However for any design matrix X, the quantity $(X^T X + \lambda I)$ is always invertible; thus there is always a unique solution for β^{ridge}. The variance of the ridge regression estimates is $Var(\hat{\beta}) = \sigma^2 W X^T X W$ Where; $W = (X^T X + \lambda I)^{-1}$ The bias of the ridge regression estimates is $Bias(\hat{\beta}) = -\lambda W \beta$;

It is shows that the total variance $\sum Var(\hat{\beta}_j)$ is a monotone decreasing sequence with respect to λ , while the total squared bias $\sum Bias^2(\hat{\beta}_j)$ is a monotone increasing sequence with respect to λ . There always exist a λ such that the MSE of β^{ridge} is less than MSE of β^{OLS} . Information criteria are a common ways of choosing among models while balancing the competing goals of fit and parsimony. In order to apply AIC or BIC to the problem of choosing λ , we will need an estimate of the degree of freedom. Recall in linear regression;

$$\hat{Y} = HY$$

$$H^{ridge} = X(X^T X + \lambda I)^{-1} X^T$$

$$df^{ridge} = \sum \frac{\lambda_i}{\lambda_i + \lambda},$$ where λ_i are the eigenvalues of $(X^T X)$ where the degree of freedom (df) is a decreasing function of λ, with $df = p$, at $\lambda = 0$ and $df = 0$, at $\lambda = \infty$.

After quantifying the (df) in ridge regression model, we can calculate AIC or BIC and use the to guide the choice of λ

.

$$AIC = n \log(RSS) + 2df$$

$$BIC = n \log(RSS) + df \log(n)$$

$$small(\lambda) \to higher(df) \to lower(RSS)$$

There always exista λ, such that the MSE of β^{ridge} is less than MSE of β^{OLS} . This is a surprising result with somewhat radical implication. Even if the model we fit is exactly correct and follows the exact distribution we specify, we can always obtain a better estimator by shrinking toward zero.

Simulation Study

The model(Yazid, 2009) considered in the simulation is

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i, \quad i=1,2,\ldots,n \quad \text{.....(5)}$$

For the simulation study, the parameters of the model in equation (5) are fixed as $\beta_0 = 0 \; and \; \beta_i = 1, i = 1,2$ The multicollinearity levels (ρ) are 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9 and 1. A Monte Carlo experiment was repeated for sample size of 10 (small), 25 (moderate) and 100 (large). At a particular choice of sample size andmulticollinearity level, a Monte-Carlo experiment was performed 1000 times.

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

$$z_1 = \frac{X_1 - \mu_1}{\sigma_1} \Rightarrow X_1 = \mu_1 + \sigma_1 z_1 \quad \text{.....(6)}$$

$$z_2 = \frac{X_2 - \mu_2}{\sigma_2} \Rightarrow X_2 = \mu_2 + \sigma_2 z_2$$

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

and

$$\beta - \frac{\sigma_{12}}{\sigma_1^2} \to \sigma_{12} - \rho \sigma_1^2 \to \sigma_2 - \frac{\beta \sigma_1}{\rho}$$

Therefore, $X_2 - \mu_2 + \frac{\beta\sigma_2}{\rho} z_1$

$$X_2 = \mu_2 + \frac{\beta\sigma_1}{\rho} z_2 \quad \text{.....(7)}$$

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

$z_{1i} \sim N(0,1), z_{2i} \sim N(0,1) and e_{ti} \sim N(0,1), i = 1,2,\ldots,1000$

The values of exogenous variables were obtained from relations (6) and (7) at different levels of multicollinearity.

RESULTS AND DISCUSSION

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) and mean squared error (MSE) criteria. The results of the analyses are presented in fig. 1-6 as follows:

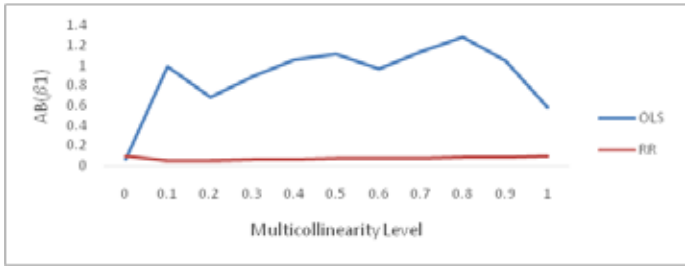

Fig 1a: Absolute Bias (AB) of β_1 at Different level of Multicollinearity with Sample size of 10

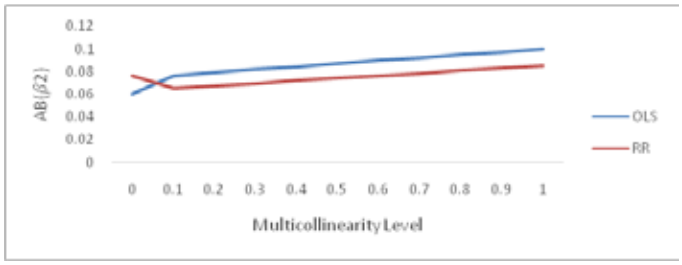

Fig 1b: Absolute Bias (AB) of β_2 at Different level of Multicollinearity with Sample size of 10

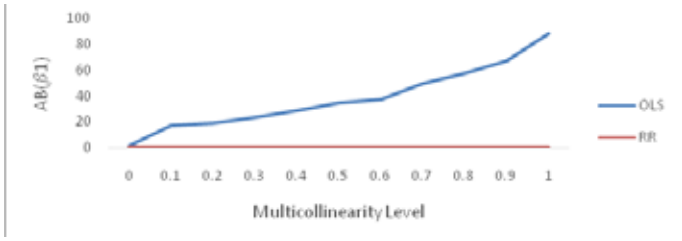

Fig 2a: Mean Squared Error (MSE) of β_1 at Different level of Multicollinearity with Sample size of 10

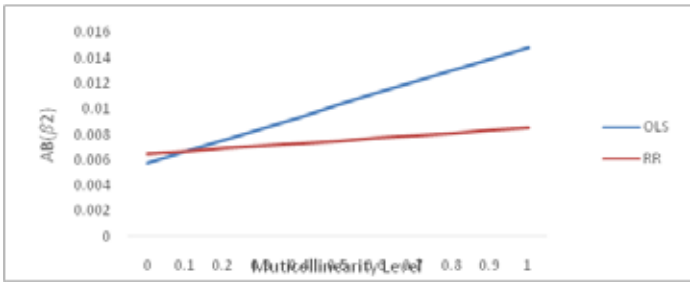

Fig 2b: Mean Squared Error (MSE) of β_2 at Different level of Multicollinearity with Sample size of 10

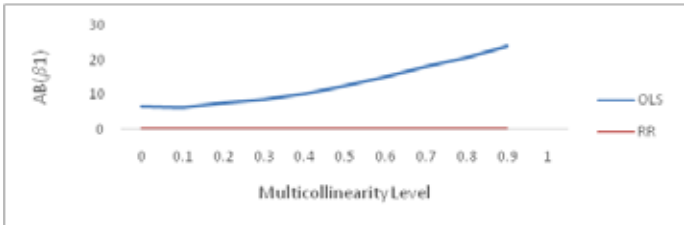

Fig 3a: Absolute Bias (AB) of β_1 at Different level of Multicollinearity with Sample size of 25

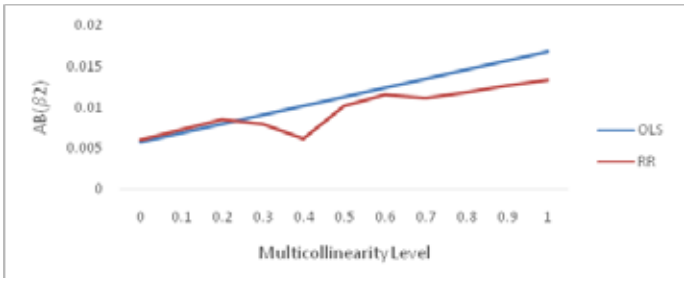

Fig 3b: Absolute Bias (AB) of β_2 at Different level of Multicollinearity with Sample size of 25

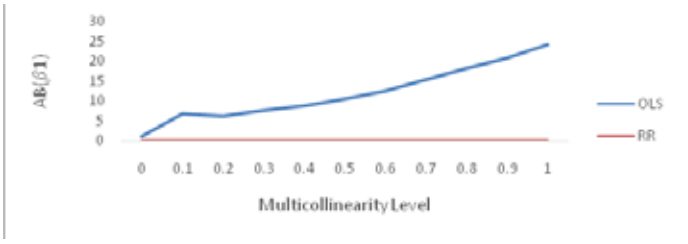

Fig 4a: Mean Squared Error (MSE) of β_1 at Different level of Multicollinearity with Sample size of 25.

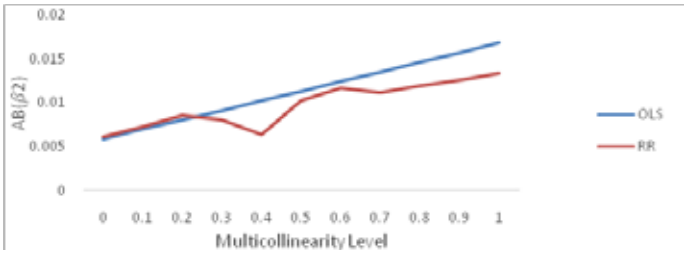

Fig 4b: Mean Squared Error (MSE) of β_2 at Different level of Multicollinearity with Sample size of 25

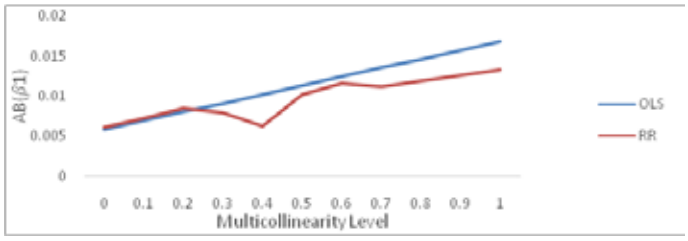

Fig 5a: Absolute Bias (AB) of β_1 at Different level of Multicollinearity with Sample size of 100

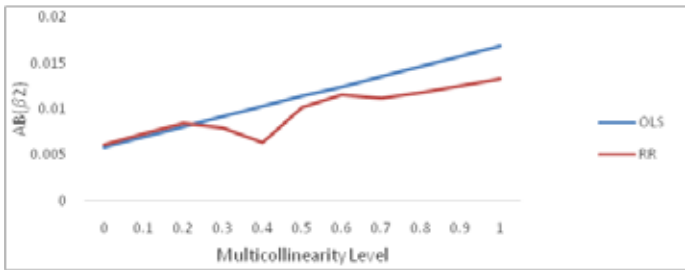

Fig 5b: Absolute Bias (AB) of β_2 at Different Level of Multicollinearity with Sample size of 100

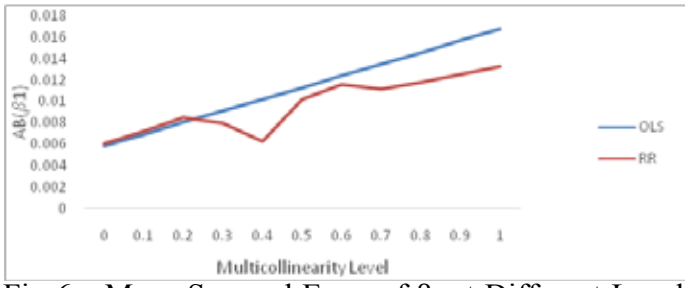

Fig 6a: Mean Squared Error of β_1 at Different Level of Multicollinearity with Sample size of 100

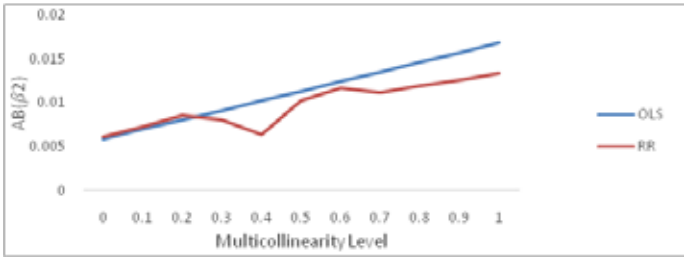

Fig 6b: Mean Squared Error of β_2 at Different Level of Multicollinearity with Sample size of 100

When there is no correlation (no Multicollinearity) of any form between the two exogenous variables in the model, results show that the parameters are best estimated with methods of OLS at different sample sizes with minimum values of both absolute bias and mean squared error. However the RR estimator outperformed OLS at the different levels of multicollinearity for small sample size (n = 25) based on both AB and MSE.

Furthermore, when multicollinearity level present in the model is not more than 0.2 for sample size of 25 and 100, the best estimators for both parameters in the model is OLS while the RR performs better than OLS from multicollinearity levels of 0.2 and above for both sample sizes. The performance of both estimators decreases with increase in multicollinearity level but increases with increase in sample size.

CONCLUSION

This study therefore shows that the RR method is the best for estimating the parameters of multiple linear regression when multicolinearity exists in the data but OLS performed better at little level of multicollinearity especially at moderate and large sample sizes. We therefore suggested that Ridge Regression should be used by the researcher when multicollinerity present in exogeneous variables while the Ordinary Least Squared can be used for data with weak correlation between the two exogenous variables of the multiple regression models for moderate and large sample size.

REFERENCES

Adnan,N; Ahmad,M. H; and Adnan,R. (2006).Comparative Study on some methods for Handling Multicollinearity problems. *Matematika*. 22(2):109-119.

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

Hoerl, A. E; Kennard, R.W. and Baldwin.(1976).Ridge Regression; Some Simulation Communication in Statistics.*Simulation and Computations. 4*: 105-123.

Khalaf, G. and Shukur.G. (2005). Choosing Ridge Parameter for Regression Problems.Communications in Statistics. *Theory and Methods*. 34:1177-1182.

Neter, J; Wasserman, W. and Kunter, M.H. (1985). Applied Linear Regression Models.2nd Ed.IRWIN Book Team. USA. 11: 77-90.

Neter J., Kunter, M.H,Nachtsein, C.J. and Wasserman,W. (1996).Applied Linear Regression Models.4th Ed.USA:IRWIN book Team.

Yazid, M .A. (2009).A Monte Carlo Comparison between Ridge and Principal Components Regression Methods. *Applied Mathematical sciences*,3(42): 2085-2098.