*Galore International Journal of Applied Sciences and Humanities*
*Vol.3; Issue: 1; Jan.-March 2019*
*Website: www.gijash.com*
*P-ISSN: 2456-8430*

*Research Paper*

# Comparison of Performance of the Least Square Regression, Principal Component Regression and Ridge Regression on Handling Multicollinearity Problem in Linear Models

Oguagbaka S. K[1], Osuji G. A[2], Aronu C. O[3]

[1]Department of Statistics, Federal Polytechnic, Oko, Anambra State, Nigeria
[2]Department of Statistics, Nnamdi Azikiwe University, Awka, Nigeria
[3]Department of Statistics, Chukwuemeka Odumegwu Ojukwu University, Anambra State

Corresponding Author: Oguagbaka S. K

## ABSTRACT

This study compared the performance of three methods on multicollinearity situation. The methods include the linear regression, the principal component regression and the ridge regression method. The methods were compared using 50 simulations and for number of independent variables p=5 and number of observation 6, 10, 20, 30, 40, 50, 60, and 100 respectively. The objectives of the study is to compare the performance of Least Squares Regression, Principal Component Regression (PCR) and Ridge Regression for handling multicollinearity problem and to determine the method that ranked best in terms of the degree of relative efficiency in overcoming the multicollinearity problem using simulated data sets. Findings of the study showed that as p is closer to n, (p=5 and n=6) the multicollinearity is very presence and evident on the R-square value for the linear regression method with is 100% and the standard error of the predicted value being zero (0). It was found that as the sample size increases, the R-square value tends to normalize. Further result showed that the Ridge regression method recorded the least R-square value while the linear regression method recorded the highest R-square value. In addition, it was found that the PCR method has the least standard deviation value across the observed sample size followed by the linear model and then the ridge regression method. This result implies that the principal component regression methods is relatively efficient for solving multicollinearity problems in linear models that the ridge regression.

***Keywords:*** *Linear Model; Multicollinearity; Standard Deviation; Relative Efficiency*

## 1.0 INTRODUCTION

Multicollinearity in practical have been found to inflate unnecessarily the standard errors of the coefficients in regression models. The increased standard errors in turn implies that the coefficients for some independent variables may be found not to be significantly far from 0 (Akinwande *et al.*,2015). However, by over amplifying the standard errors, multicollinearity makes some the variables in the regression model statistically insignificant especially when they should be significant. This has posed serious threat to the usefulness of the regression model in making acceptable estimation.

It is known that the problem of multicollinearity is present in the data set where the number of variables is high compared to the number of observations (i.e. when p >n). Also, it was observed from the review of literature that there is limited literature on handling of multicollinearity problems associated with the number of independent variables (p) being very close to the number of observation (n). Hence, the motivation to examine the performance of the least square method, the principal component regression and the ridge regression on situation where the number of observation (n) is at least one unit greater than the number of predictor variables.

The aim of this study is to examine the performance of the least square regression analysis, principal component regression analysis, and the ridge regression analysis on multicollinearity situation with the following specific objectives: to compare the use of Least Squares Regression, Principal Component Regression and Ridge Regression for handling multicollinearity problem and to determine the method that ranked best in terms of the degree of relative efficiency in overcoming the multicollinearity problem using simulated data from the standard normal distribution.

## 2. LITERATURE REVIEW

According to Mason and Perreault (1991), numerous approaches have been proposed for coping with collinearity-none entirely satisfactory. Just like procedures for detection, the procedures for coping with collinearity vary in level of sophistication. Their study reviewed in brief several of the most commonly used approaches for coping with collinearity. The authors argued that one of the simplest responses to coping with collinearity is to drop one or more of the collinear variables. This approach may eliminate the collinearity challenge, but it prompts new complications. In the real sense, unless the true coefficient(s) of the dropped variable(s) is zero, the model will be wrongly specified, resulting in biased estimates of some coefficients which might be costly. Second, dropping variables makes it difficult to identify the relative usefulness of the predictor variables. Even when one disregards these limitations, the practical problem of deciding which variable to drop remains unsolved.

Graham (2003) in his study opined that although multiple regression is commonly used in testing the individual effects of many explanatory variables on a continuous response, the inherent collinearity (multicollinearity) of confounded explanatory variables limits analyses and threatens their statistical and inferential interpretation. The study employed numerical simulations, to quantify the impact of multi-collinearity on ecological multiple regression and found that even low levels of collinearity bias analyses such as correlation coefficient value of $r \geq 0.28$ or $r^2 \geq 0.08$, can result to any of the following, inaccurate model parameterization, decreased statistical power, or exclusion of significant predictor variables during model creation.

Vatcheva *et al.* (2016) reviewed epidemiological literature in Public Medicine (PubMed) from January 2004 to December 2013. From the findings of their review, they suggested the need for more attention in identifying and minimizing the effect of multicollinearity in analysis involving data from epidemiologic studies. The study employed data generated from simulation and real life data from the Cameron County Hispanic Cohort to demonstrate the adverse effects of multicollinearity in the regression analysis. They advised researchers to consider the diagnostic for multicollinearity as one of the steps in regression analysis.

According to Mela and Kopalle (2002), the problem of collinearity in empirical research is among the most endemic concerns raised by researchers in various fields especially management sciences. A recent search in EconLit revealed 154 studies discussing collinearity or multicollinearity in their abstracts. A similar full text search of Applied Economics (using Infotrac) yielded 220 articles since 1991 (Mela and Kopalle, 2002). Various econometric references have indicated that collinearity increases estimates of parameter variance, yields high coefficient of determination in the face of low parameter significance, and results in parameters with incorrect signs and implausible magnitudes (Belsley *et al.*, 1980; Kmenta, 1986).

Wentzell and Montoto (2003) compared the principal components regression (PCR) and the partial least squares regression (PLS) using simulation studies of complex chemical mixtures which contain a large number of components. The findings of their study showed how the prediction errors and number of latent variables (NLV) used vary with the relative abundance of mixture components. Simulation parameters varied include the distribution of mean concentrations, spectral correlation, noise level, number of mixture components, number of calibration samples, and the maximum number of latent variables available. They found that in all cases, except when artificial constraints were placed on the number of latent variables retained, there exist no significant differences in the prediction errors reported by PCR and PLS. Also, they observed that the PLS almost always required fewer latent variables than PCR, but did not influence predictive ability.

Chopra *et al*. (2013) employed two regression methods namely; the traditional regression method and the ridge regression method for the prediction of the estimate of the response variable. They varied the values of the regression coefficients drastically such that negative coefficients were transformed into positive and positive coefficients transformed into negative when regression analysis was employed and data reduced or raised. The findings of their study showed that the traditional method did not prove to be credible for forecasting the estimate of the dependent variable (compressive strength of concrete). However, both in the case of reduction and augmentation of data, there exist frequent minimum effect which has no or negligible impact on the coefficients when performing the Ridge regression method.

## 3.0 RESEARCH METHODOLOGY

### 3.1 Method of Data Collection

The source of data used for this study is includes simulation from standard normal distribution will be used for p=5 with n= 6, 10, 20, 30, 40, 50, 60, and 100 respectively.

### 3.2 The Least Square Regression

The least square method of estimating regression parameters aims at generating estimators in such a way that the sum of squares of the error is minimized.

Suppose,

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

where X is an n x (k+1) matrix of full rank,
β is a (k+1) x 1 vector of unknown parameters,
and ε is an n x 1 random vector with mean 0 and variance $\sigma^2$I.
The least squares estimator for β is denoted by b and is given by

$$\hat{b} = (X'X)^{-1} X'y \quad (2)$$

To solve for b, we shall multiply both sides by $(X'X)^{-1}$ to obtain the least square estimators as

$$\text{b} = (X'X)^{-1} X'y \ .$$

The covariance matrix of $\hat{\text{b}}$ is equal to

$$\text{COV}(\hat{\text{b}}) = \sigma^2 \left(X'X\right)^{-1} \quad (3)$$

Equation (3) can also be written as

$$\text{COV}(\hat{\text{b}}) = \sigma^2 \sum_{i=k}^{K} \text{p}_k \left(\frac{1}{\lambda_k}\right) \text{p}'_k \quad (4)$$

where the p's are the eigenvectors of $\text{X}'\text{X}$ and the λ's are the corresponding eigenvalues.

### 3.3 The Principal Components Regression

Principal Components 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, principal components regression reduces the standard errors. It is hoped that the net effect will be to give more reliable estimates.

The least square regression parameters are estimated by equation (2).

Since the variables are standardized, $X'X = R$, where R is the correlation matrix of independent variables.

To perform principal components (PC) regression, the independent variables to their principal components are transformed. Mathematically, we write

$X'X = PDP' = Z'Z$ (3)

where D is a diagonal matrix of the eigenvalues of $X'X$,

P is the eigenvector matrix of $X'X$, and

Z is a data matrix (similar in structure to X) made up of the principal components.

P is orthogonal so that $P'P = I$.

A new variable Z will be created as weighted averages of the original variables X. Since these new variables are principal components, their correlations with each other are all zero. Hence, for three independent variables, we shall begin with variables $X_1$, $X_2$, and $X_3$, and end up with $Z_1$, $Z_2$, and $Z_3$.

Severe multicollinearity will be detected as very small eigenvalues. To rid the data of the multicollinearity, we omit the components (the z's) associated with small eigenvalues. Usually, only one or two relatively small eigenvalues will be obtained. For example, if only one small eigenvalue were detected on a problem with three independent variables, we would omit Z3 (the third principal component). When we regress Y on Z1 and Z2, multicollinearity is no longer a problem. We can then transform our results back to the X scale to obtain estimates of B. These estimates will be biased, but the size of this bias is expected to be more than compensated for by the decrease in variance. This implies that the mean squared error of these estimates is expected to be less than that for least squares. Mathematically, the estimation formula becomes

$\hat{A} = \left(Z'Z\right)^{-1} Z'Y = D^{-1} Z'Y$ (4)

because of the special nature of principal components. Equation (4) resembles equation (2) but applied to a different set of independent variables, Z. Note that the two sets of regression coefficients, A and B, are related using the expression given as:

$A = P'B$ (5)

and

$B = PA$ (6)

(Recall that P is orthogonal so that $P'P = I$, thus by multiplying both sides of equation (5) by P will yield equation (6)).

To perform the principal component analysis, the covariance matrix is factored out using the spectral decomposition theorem which is expressed as

$\sum = \lambda_1 e_1 e_1' + \lambda_2 e_2 e_2' + \Lambda + \lambda_p e_p e_p'$ (7)

where $(\lambda_i, e_i)$ is the eigenvalue-eigenvector pair of $\sum$ and $\lambda_1 \geq \Lambda \geq \lambda_p \geq 0$.

This fits the covariance structure of the factor analysis model having as many factors as variables (m = p) and having a specific variance of $\psi_i = 0$ for all i.

$\sum_{p\times p} = L_{p\times p} L'_{p\times p} + 0_{p\times p} = LL'$ (8)

The model which includes the specific factors and their variances are taken to be the diagonal elements of $\sum - LL'$ and the approximation is given as

$\sum = LL' + \psi$ (9)

where,

$$\psi = \begin{pmatrix} \psi_1 & 0 & \Lambda & 0 \\ 0 & \psi_2 & \Lambda & 0 \\ M & M & M & M \\ 0 & 0 & \Lambda & \psi_p \end{pmatrix}$$

The number of factors to be retained is similar to the number of positive eigen values of the correlation matrix.

The principal component regression estimator can be expressed as

$\hat{Y}_k = \left(Z_k' Z_k\right)^{-1} Z_k' y$ (10)

where,

$$Z_k = XV_k = [Xv_1, \Lambda , Xv_p]$$

The final principal component regression estimator of β based on using the first k principal component is given as

$$\hat{\beta}_k = V_k \hat{Y}_k \quad (11)$$

In addition, the assumptions for the PC regression are the same as those used in regular multiple regression: linearity, constant variance (no outliers), and independence. However, since PC regression does not provide confidence limits, normality need not be assumed.

### 3.2.3 The Ridge Regression

One of the goals of ridge regression is to produce a regression equation with stable coefficients. The coefficients are stable in the sense that they are not affected by slight variations in the estimation data. The ridge regression approach is an attempt to construct an alternative estimator that has a smaller total mean square error value.

Recall that the least square regression parameters are estimated by equation (2).

To stabilize the coefficients of the regression model, $kI$ is added to $X'X$ in equation (2) as was proposed by Hoerl (1975). Hoerl (1975) named this method ridge regression because of its similarity to ridge analysis used in his earlier work to study second-order response surfaces in many variables.

Hence, the ridge regression is defined by

$$\hat{b}(k) = \left(Z'Z + kI\right)^{-1} Z'y \quad (12)$$

where k is the bias parameter and $Z_j$ is obtained by transforming the original predictor variable $X_j$ by

$$Z_{ij} = \frac{(x_{ij} - \bar{x}_j)}{\sqrt{(x_{ij} - \bar{x}_j)^2}} \quad (13)$$

As k increases from zero, bias of the estimates increases. On the other hand, the total variance (the sum of the variances of the estimated regression coefficients), is

$$\text{Total variance}(k) = \sum_{j=1}^{p} \text{Var}(\hat{\beta}_j) = \sigma^2 \sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j + k)^2}$$

(14)

where, $\lambda_j$ is the corresponding eigenvalues Equation (3.9) is a decreasing function of k. This shows the effect of the ridge parameter on the total variance of the ridge estimates of the regression coefficients. Substituting k = 0 in (3.9), we obtain

$$\text{Total variance}(k) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j} \quad (15)$$

This shows the effect of small eigenvalue on the total variance of the Ordinary Least Square (OLS) estimates of the regression coefficients. As k continues to increase without bound, the regression estimates all tend toward zero.

The idea of ridge regression is to pick a value of k for which the reduction in total variance is not exceeded by the increase in bias. It has been shown that there is a positive value of k for which the ridge estimates will be stable with respect to small changes in the estimation data (Hoerl and Kennard, 1970).

#### 3.2.3.1 Fixed Point Method of Determining k

The fixed point method of estimating k was suggested by Hoerl *et al.*, (1975). The method is expressed mathematically as:

$$k = \frac{p\,\sigma^2(0)}{\sum_{j=1}^{p} (\hat{\beta}_j(0))^2} \quad (16)$$

where $\hat{\beta}_1(0), \Lambda , \hat{\beta}_p(0)$ are the least squares estimates of $\hat{\beta}_1, \Lambda , \hat{\beta}_p$ when the model in (2) is fitted to the data (i.e., when k = 0), and $\sigma^2(0)$ is the corresponding residual mean square.

## 4. DATA ANALYSIS AND RESULTS

### 4.1 Comparison of adequacy measure of the Methods

This section presents the coefficient of determination measure of the three methods for the various situations considered in this study.

**Table 1: Summary Result of R-square Value**

| Sample Size | LM | PCR | RIDGE |
|---|---|---|---|
| 6 | 100 | 79.85794 | 75.33747 |
| 10 | 85.79662 | 83.7748 | 36.99988 |
| 20 | 83.84568 | 83.42494 | 37.21827 |
| 30 | 82.64182 | 80.05734 | 37.1226 |
| 40 | 80.22684 | 79.19647 | 53.59439 |
| 50 | 79.59645 | 73.81498 | 58.5788 |
| 60 | 78.97764 | 71.54733 | 46.41893 |
| 100 | 71.21008 | 69.49055 | 38.39548 |

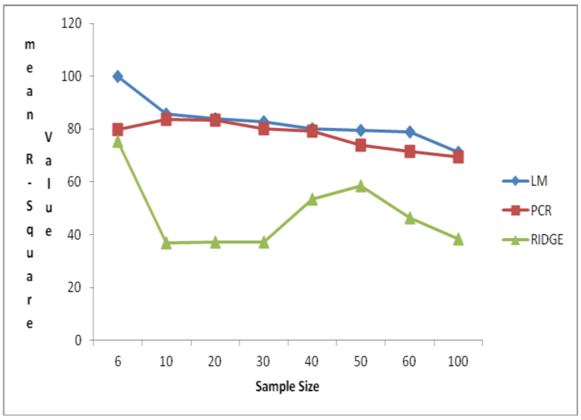

**Figure 1: Distribution of mean R-square Value for p=5**

The result presented in table 1 shows the summary of the mean R-square values obtained for 50 simulation for the observed sample sizes. This result is employed to plot a line graph of the methods in figure 1. The result revealed that the Ridge regression method recorded the least R-square value while the linear regression method recorded the highest R-square value.

## 4.2 Comparison of relative efficiency of the Methods

This section presents the relative efficiency of the three methods for the various situations considered in this study.

**Table 2: Summary Result of Relative Efficiency of the predicted value**

| Sample Size | LM | PCR | RIDGE |
|---|---|---|---|
| 6 | 0 | 0.9464 | 0.8081 |
| 10 | 1.0018 | 0.7663 | 0.771 |
| 20 | 1.1098 | 0.6372 | 0.8396 |
| 30 | 1.0296 | 0.8606 | 0.9334 |
| 40 | 1.0369 | 0.8777 | 0.9966 |
| 50 | 0.9503 | 0.8457 | 0.9403 |
| 60 | 0.8677 | 0.7363 | 0.8203 |
| 100 | 0.7978 | 0.7183 | 0.7455 |
| Mean (SD) | 0.8492 | 0.7985 | 0.8569 |

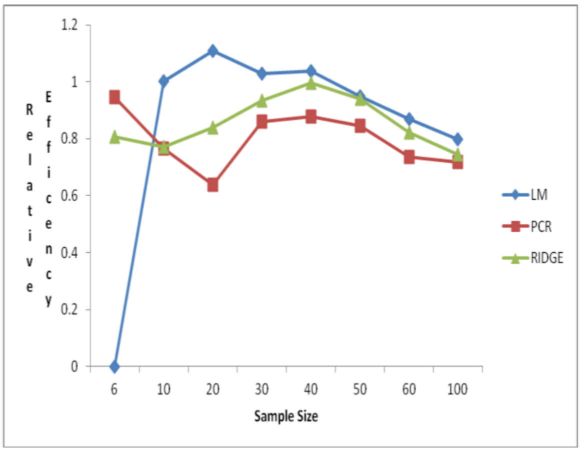

**Figure 2: Distribution of relative efficiency of the predicted value**

The result presented in table 2 shows the summary of the PCR method has the least standard deviation value across the observed sample size with an average of 0.7985, followed by the linear model with a value of 0.8492 and the ridge regression method with a value of 0.8569. This result is employed to plot a line graph of the methods in figure 2. The result revealed that the principal component regression methods is relatively efficient for solving multicollinearity problems in linear models that the ridge regression.

## 5. CONCLUSION

This study compared the performance of three methods on multicollinearity situation. The methods include the linear regression, the principal component regression and the ridge regression method. The methods were compared using 50 simulations and for number of independent variables p=5 and number of observation 6, 10, 20, 30, 40, 50, 60, and 100 respectively. Findings of the study showed that as p is closer to n (p=5 and n=6) the multicollinearity is very presence and evident on the R-square value for the linear regression method with is 100% and the standard error of the predicted value being zero (0).

It was found that as the sample size increases, the R-square value tends to normalize. Further result showed that the Ridge regression method recorded the least R-square value while the linear regression method recorded the highest R-square value.

In addition, it was found that the PCR method has the least standard deviation value across the observed sample size followed by the linear model and then the ridge regression method. This result implies that the principal component regression methods is relatively efficient for solving multicollinearity problems in linear models that the ridge regression. This study considered number of independent variables p=5, we recommend using p>5 as area for further research.

## REFERENCES

- Akinwande, M. O., Dikko, H. G. and Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. *Open Journal of Statistics*, 5:754-767
- Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics -Identifying Influential Data and Sources of Collinearity*. John Wiley and Sons, New York.
- Chopra, P., Sharma, R. K. and Kumar, M. (2013). Ridge Regression for prediction of compressive Strength of Concrete. *International Journal of Innovation in Engineering Technology*, 2: 106-111.
- Hoerl, A. and Kennard, R. and Baldwin, K. (1975). Ridge Regression: Some Simulations, Communication in Statistics - Theory and Methods 4: 105–123.
- Hoerl, A. and Kennard, R. (1970). Ridge Regression: Biased Estimation for Non orthogonal Problems, *Technometric*, 12: 55 – 67.
- Kmenta, J. (1986). *Elements of Econometrics(2nd edn)*. Macmillan Publishing Company, New York.
- Mason, C. H. and Perreault, W. D. (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research*, 28(3): 268-280
- Mela, C. F. and Kopalle, P. K. (2002). The impact of Collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*, 34: 667-677.
- Vatcheva, K. P., Lee, M., McCormick, J. B. and Rahbar, M. H. (2016). Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiology*, 6(2): 227- 235.
- Wentzell, P. D. and Montoto, L. V. (2003). Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems* 65: 257–279.

How to cite this article: Oguagbaka SK, Osuji GA, Aronu CO. Comparison of performance of the least square regression, principal component regression and ridge regression on handling multicollinearity problem in linear models. Galore International Journal of Applied Sciences & Humanities. 2019; 3(1): 1-7.

******