COMPARATIVE ANALYSIS OF THE EFFICIENCIES ON METHODS OF HANDLING MULTICOLLINEARITY IN REGRESSION ANALYSIS

E. O. Ogunjobi [1] , D. A. Agunbiade [2] , O. A. Ayansola [3]

[1] Department of Mathematics and Statistics, The Ibarapa Polytechnic, Eruwa Oyo State, Nigeria
[2] Department of Mathematical Sciences, Olabisi Onabanjo University, Ago Iwoye Ogun State, Nigeria
[3] Department of Mathematics and Statistics, The Polytechnic Ibadan, Ibadan Oyo State, Nigeria

Corresponding Author: E. O. Ogunjobi, eyitayo_ogunjobi@yahoo.com

ABSTRACT: This paper considers three different approaches of handling multicollinearity in regression analysis using economic data. These techniques were applied to study determinants of money supply of the sector of the economy. The goal is to determine which of the economic variables included in the factors that influence money supply (either in the broad 'M1' or the narrow 'M2'sense) is not actually contributing to the effect of money supply by monetary authority of the sector of the economy. A comparative analysis of the three methods using the adjusted R^2 , Mean Square Error and Root Mean Error, as the statistics criteria, revealed that the result obtained from the Ridge Regression gave the outstanding performance as compared with the other two techniques. On the overall, it was found that the Ridge Regression performed best, followed by Principal Component Regression, while Latent Root Regression performed least.

KEYWORDS: Multicollinearity, Ordinary Least Squares, Principal Component Regression, Latent Root Regression and Ridge Regression.

1. INTRODUCTION

When two or more independent variables in a regression model are interrelated (correlated), one of them is bound to be redundant. Multicollinearity exists when independent variables in a regression model are related. When this situation exists, the Ordinary Least Squares (OLS) method becomes inappropriate because the regression coefficients are inefficient (with large variances), thus result to biased regression methods. Gujarati and Porter ([GP09]) gave an exact relation for k - variable regression as

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots\ldots + \lambda_k X_k = 0$$

Where λ_1, λ_2 ..... λ_k are constants such that not all of them are zero.

The presence of multicollinearity in least squares regression can cause large variances of parameter estimates which mean that the estimates of the parameters tend to be less precise. As a result, Johnston ([Joh03]) argued that the model will have insignificant parameters and wide confidence interval.

Therefore, the more the multicollinearity, the less interpretable are the parameters. Multicollinearity among the regressor variables is a serious problem which may dramatically impair the usefulness of a regression model. Since the regressors are the columns of the X – matrix, an exact linear dependent would result in a singular $X'X$. When their exact dependencies between columns of the X matrix, that is when one or more columns can be exactly expressed as a linear combination of other columns, det $(X'X) = 0$ .This may either mean that the model is over specified i.e. more parameters are postulated than are needed to express the data or the data are not adequate to estimate the model set down.

Correlation matrix, Variance inflation factor (VIF) are used to detect multicollinearity. ([MS70]) is of the opinion that VIF greater than 10 indicates multicollinearity.

We can also remedy data from multicollinearity situation by the use of "ratio of parameter" (by transformation).

After this transformation, we then go ahead using least square method to estimate the parameter.

Some controversies surrounding whether the regressor and the response should be centered and scaled so the $X'X$ and $X'Y$ are in correlation form which result in an artificial removal of the intercept from the model. The intercept in the ridge regression model is estimated by y. According Marquardt and Snee ([MS75]), centering tend to minimize an ill conditioning when fitting polynomials. Belsley, Kuh and Welseh ([WBK80]), do not support centering of the regressor too, so that the role of the in any near linear dependencies may be diagnosed. Centering and scaling allows the analyst to think of the parameter estimates as standardized regression coefficient, which is often appealing. Furthermore centering and scaling remove ill conditioning thereby reducing variance inflation in the parameter estimates. In the study of the predictor variables, the mean is subtracted from each of the variables

(centering) and the centered variables are divided by its own standard deviation (scaling). Centering removes non essential ill – conditioning, also removes the correlation between the constant term and all the linear terms. In addition a quadratic model centering reduces and in certain situations completely removes the correlation between the linear and the quadratic terms. Scaling expresses the equation in a form that lends itself to more straight forward interpretation and use.

A number of Monte Carol simulations have been conducted to examine the effectiveness of biased estimator and to attempt to determine which processes perform best. Norliza et al. ([N+07]), compared Ridge Regression (RR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR), while no single procedure emerges from these studies as the best overall, there is considerable evidence indicating the superiority of biased estimation techniques to the Least Squares especially. Also Al-Hassan ([AA09]), compared the Ridge Regression and Principal Component based on the simulation study, he found that Ridge Regression method performs better than Principal Component method, El-Dereny and Rashwan ([ER11]), worked on the Ridge Regression and Ordinary Least Squares based on their results they found that Ridge Regression are better than Ordinary Least Squares method if the multicollinearity exists in the set of data.

2. MATERIALS AND METHODS

2.1 Principal Component

This study examines the works of some researcher notably Thiel ([Thi71]), Morrison ([Mor76]), Johnston ([Joh03]), Bjorkstrom and Sundberg ([BS99]) who developed the theoretical description of Principal Component Regression analysis. Let X be a p variate observation vector distributed normally with mean vector μ and covariance matrix Σ i.e $X \sim N_p(\mu, \Sigma)$. If p is large, the random vector may be difficult to handle. We may therefore reduce p to a manageable size, say k, where $k < p$ considering k – linear combination of X say $y_i = a_i X$, i = 1, 2,……k. Suppose there are $N \times P$ variate observations given by the matrix:

$$X = \begin{pmatrix} X_{11} & X_{12}\ldots\ldots X_{1P} \\ X_{21} & X_{22}\ldots\ldots X_{2P} \\ \ldots\ldots\ldots\ldots\ldots \\ X_{N1} & X_{N2}\ldots\ldots\ldots X_{NP} \end{pmatrix} = \begin{pmatrix} X'_1 \\ X'_2 \\ \ldots. \\ X'_N \end{pmatrix}$$

Suppose the transforming coefficient (which is not yet determined at this point) is:

$$a_1 = \begin{pmatrix} a_{11} \\ a_{12} \\ . \\ a_{1p} \end{pmatrix} \quad a_2 = \begin{pmatrix} a_{21} \\ a_{22} \\ . \\ a_{2p} \end{pmatrix} \ldots\ldots n_k = \begin{pmatrix} a_{k1} \\ a_{k2} \\ . \\ a_{kp} \end{pmatrix}$$

$$y_{11} = Xa_1, \quad y_{21} = Xa_2, \ldots\ldots, y_{N1} = Xa_{1p}$$

Where

$$y_{11} = \begin{pmatrix} X_{11} & X_{12}\ldots\ldots X_{1P} \\ X_{21} & X_{22}\ldots\ldots X_{2P} \\ \ldots\ldots\ldots\ldots\ldots \\ X_{N1} & X_{N2}\ldots\ldots\ldots X_{NP} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ . \\ a_{1p} \end{pmatrix}$$

$$y_{21} = \begin{pmatrix} X_{11} & X_{12}\ldots\ldots X_{1P} \\ X_{21} & X_{22}\ldots\ldots X_{2P} \\ \ldots\ldots\ldots\ldots\ldots \\ X_{N1} & X_{N2}\ldots\ldots\ldots X_{NP} \end{pmatrix} \begin{pmatrix} a_{21} \\ a_{22} \\ . \\ a_{2p} \end{pmatrix}$$

Hence we shall obtain:

$$\begin{aligned} y_{11} &= a_{11}X_{11} + a_{12}X_{12} + \ldots + a_{1p}X_{1p} \\ y_{21} &= a_{11}X_{21} + a_{12}X_{22} + \ldots + a_{2p}X_{2p} \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{N1} &= a_{11}X_{N1} + a_{12}X_{N2} + \ldots + a_{N}pX_{Np} \end{aligned}$$

This implies that if we use the data matrix X on a_1, we shall obtain N values

$y_{11}, y_{21}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots y_{N1.}$

Similarly, using the data matrix X on a_2 we shall obtain another N values,

$y_{12}, y_{22}, \ldots\ldots\ldots\ldots\ldots\ldots y_{N2}$ and so on.

Therefore, if we use the data matrix on the vector a_1, $a_2, \ldots a_k$. We shall obtain the transformed matrix Y which is defined by:

$$Y = \begin{pmatrix} y_{11} & y_{12}\ldots\ldots y_{1k} \\ y_{21} & y_{22}\ldots\ldots y_{2k} \\ \ldots\ldots\ldots\ldots\ldots \\ y_{N1} & y_{N2}\ldots\ldots\ldots y_{Nk} \end{pmatrix}$$

Which of course consist of N observations as before but each observation vector is now a k variate ($K \leq p$) and each element of the ith k- vector is a linear combination of the p element of the original i^{th} p – variate vector.

Carrying out this transformation, we would wish to minimize the "loss of information" in the original data. If $p < k$ (i.e minimized the sum of square of all N_p error discrepancies).This is therefore done by choosing a_1 such that the variance of yi is minimized. That is we aim to incorporate as much of

variation within the *p* variate structure into *yi* variate. The second vector a_2 is chosen so that the resulting y_2 is dependent of y_1 and incorporate as much as the remaining variation. In such a way, a total of *k* independent linear combinations can be chosen which extract as much variation as possible.
Suppose now that the random variable X_1, X_2, X_3……X_p have a *p* variate normal distribution with mean vector μ and co-variance matrix Σ. Let $\boldsymbol{\lambda i}$: i = 1, 2, ……p be the characteristic root (Latent Root) of Σ and let $\boldsymbol{\lambda_i}$'s be ordered so that $\boldsymbol{\lambda i} \geq \boldsymbol{\lambda i} = 1$. Let *a*i be the associated normalizes eigenvector so that *ai'ai = 1* (restriction). Thus a_1 is the normalized eigenvector corresponding to the largest eigenvalue λ1, and a_2 is the normalized eigenvector corresponding to the second largest eigenvalue (characteristic root) $\boldsymbol{\lambda}_2$ and so on. Then $yi = a'_1X$ is then linear compound.

$$a_{11}X_1 + a_{12}X_2 + \ldots\ldots\ldots\ldots\ldots + a_{1p}\ X_p$$

of the responses is called the first Principal Component of *X* and in general, *y1 = a'iX* is called the *ith* Principal Component of *X*.
The above discussion is made clearer with this illustration using matrix notations. Assume we have *n* observations on *k* variables arranged in an *n x k* matrix of *X* (as shown above). If we have one single variable having n - values arranged in a column vector *p* (*p* is not yet determined). Assume it were, if all the variables have proportionally each column of *X* equal to some scalar multiple of *p* hence we say *X = pa'* where a' is the *k* element row vector consisting of the scalar multiple. One of each column of *X*.*X = pa'* will in general hold' exactly hence, there will be a non – zero error matrix *X – pa'* for all vector *p* and a .We now select these vectors such that the sum of square of all *kn* errors (loss of information) is minimized. The sum of squares of all elements a_{ij}^2 of the *m x n* . A matrix can be written as the trace of A'A.

$$tr\ A'A = \sum_{i=1}^{m} \sum_{j=1}^{n} a^2 ij$$

We therefore seek to minimize:

$$tr(X - pa')'(X - pa') = tr\,(X'X) - tr\,(ap'X) - tr(X'pa) + tr(ap'ap\ ')\ but$$
$$tr\,(ap\ 'X) = tr(p'pa'a) = p'pa'a = a'a$$

for uniqueness we have to impose the condition p'p = 1.
This implies that

$$tr\,(X - pa')\ '(X - pa') = tr\,(X'X) - 2p'Xa + a'\,a \qquad (1)$$

Now to minimize (1) above we differentiate it partially with respect to a given p and equate the result to zero.
Let $\tau = tr(X - pa')'(x - pa')$

$$\therefore \frac{\delta\tau}{\delta a} = -2p'X + 2a,$$

if

$$\frac{\delta\tau}{\delta a} = 0$$

then,

$$a - X'p = 0 \quad \Rightarrow \quad a - X'p \qquad (2)$$

Substituting (2) into (1)

$$tr\,(X'X) - 2p'\,X'X\,p + p'XX'p = tr(X'X) - p'XX'p$$

This shows that we maximize *p'X'Xp* the variance of *y* for variation in *p* subject to our restriction p'p = 1. To do this, we employ the lagrangian expression

$$\phi = pXX'p - \boldsymbol{\lambda}(p'p - 1)$$

where $\boldsymbol{\lambda}$ is the lagrangian multiplier.
Differentiating with respect to *p*:

$$\frac{\delta\phi}{\delta p} = 2pXX' - 2\boldsymbol{\lambda}p$$

Setting $\frac{\delta\phi}{\delta p} = 0$,

then,

$$pXX' - \boldsymbol{\lambda}\,p = 0$$
$$pXX' - \boldsymbol{\lambda}\,p \text{ or } (XX' - \boldsymbol{\lambda}I) = 0 \qquad (3)$$

p is the characteristic vector of *n x n* positive semi-definite matrix *XX'* corresponding to root $\boldsymbol{\lambda}$, (i.e. *p* is the eigenvector of X'X corresponding to root $\boldsymbol{\lambda}$).
The largest root *XX'* would be taken since we are maximizing *p' XX'p*. Therefore, we pre – multiply (3) by *p'*

$$P'(XX' - \boldsymbol{\lambda}I)\,p = 0$$
$$= p'\,XX'p - \boldsymbol{\lambda}p'p = \boldsymbol{\lambda}$$

Also pre – multiply (3) by *X'* we have

$$X'(XX' - \boldsymbol{\lambda}I)p = 0 \ \Rightarrow (XX' - \boldsymbol{\lambda}I)X'\,p = 0$$

Recall from (2) *X'p = a:* $(XX' - \boldsymbol{\lambda}I)\,a = 0$

The coefficient vector a is hence the characteristics vector of XX' matrix which corresponds to the largest root (though un-normalized to have unit length).
From (3) $\lambda p = X(X'P) = Xa$

$$\therefore \quad p = 1/\lambda \; Xa \qquad (3a)$$

The vector p thus derived is referred to as the first principal component of the k – variable represented in X. It gives the best linear description of the X-column in the least square sense.

2.2 Latent Root

Webster et al. ([W+76]), Latent root regression follows the same principle as the Principal Component Regression. In fact, it can be rightly referred to as an extension of Principal Component for examining alternative prediction equation and for the elimination of predictor variables. It was first proposed by Webster, Gunst and Mason ([W+76]) – Latent Root regression attempts to identify and eliminate multicollinearity. It will be noted that it is reduced to the Least Squares when no terms are actually deleted from the original data. Gunst et al. ([W+76]) and Gunst and Mason ([GM77]), indicate that Latent Root Regression may provide considerable improvement in the Mean Square Error (MSE) over Least Squares. Gunst and Mason ([GM80]), points out that it can produce regression coefficients that are very similar to those found by Principal Component Regression, particularly when there are only one or two strong multicollinearity in X data.
Webster and his co-workers augmented the data matrix of the centered and scaled predictor variables with the centered and scaled response variable, placing the latter first in order to provide $Z^{\prime} = (y, Z)$ where Z is the centered and scaled " x matrix, "

$$y = (Y - I\bar{Y}\ )/S^{1/2}\ yy$$

and

$$\text{Syy} = \Sigma(Y - \bar{Y})^2$$

It follows that Z' Z is the augmented correlation matrix. As in the Principal Component analysis method, the Latent Roots or characteristics roots or eigenvalues, and their corresponding Latent vectors are calculated, but here the first element (the "Y coefficient") of each of the latent vectors is used as a measure of the predictability of the response by the latent vectors. The larger the size of the first element of the latent vector the more useful is that vector in predicting the response and vice versa.
The presence of small Latent Root indicates potential linear dependence among the predictor variables. The smaller the Latent Root the more pronounced is the dependence. The Least Squares equation for a given model is the best (in Least Squares sense) linear combination of all these latent vectors. By dropping out latent vectors whose latent roots and corresponding first element of the latent vectors are small, a modified Least Squares estimation equation is obtained. This modified procedure leads to biased estimators.
Expressing theoretically the working of this procedure, we first obtain the augmented correlation matrix from the original data.

$$\begin{array}{c} \\ Y \\ Z_1 \\ Z_2 \\ . \\ Z_n \end{array} \begin{array}{c} \begin{array}{ccccc} Y & Z_1 & Z_2 & \ldots\ldots & Z_n \end{array} \\ \begin{pmatrix} 1 & r_{yz1} & r_{yz2} & \ldots\ldots & r_{yzn} \\ & 1 & r_{z1z2} & \ldots\ldots & r_{z1zn} \\ & & 1 & \ldots\ldots & \\ & & & \ldots\ldots\ldots\ldots & \\ & & & & 1 \end{pmatrix} \end{array}$$

It will be a symmetric matrix where all the diagonal elements are 1 (unit) since each is the correlation between Zj Z_j i.e $r_{jj} = 1$
The Latent Root X_j of the augmented matrix and their corresponding latent vector rj are next to be determined. In the display below a Latent Root λj is at the left of each row while the remainder of the row being the vector corresponding to that Latent Root.

Latent root	Y	Z_1	Z_2	.	.
.	Z_j				
λ_j	γ_{0j}	γ_{1j}	γ_{2j}		
	γ_{nj}				
γ_p	.	.	.	.	.
.	.				
γ_{p-1}	.				
.	.				
γ_0	γ_{00}	γ_{01}	γ_{02}	.	.
.	γ_{0p}				

Next is to examine the Latent Roots λ_j. and their corresponding γ_{0j} values. If for any j, both these values are small, near singularity in that data and poor response predictability are indicated. The guidelines for smallness proposed by Webster et al. ([W+74]), are $\lambda_j \leq 0.05$ and $\lambda_{0j} \leq 0.10$ such vector should be removed from consideration.
Here, we now perform the estimation procedure. Firstly, we decide via step three which vector we wish to retain. Then the vector of the modified Least Squares coefficient is given by the formula.

$$b^{\prime} = \begin{pmatrix} b_1^{\prime} \\ b_2^{\prime} \\ . \\ . \\ b_r^{\prime} \end{pmatrix} = C\Sigma^{\prime}_j \; \gamma_{0j} \; \lambda j^{-1} \begin{pmatrix} \gamma_{1j} \\ \gamma_{2j} \\ . \\ . \\ \gamma_{rj} \end{pmatrix}$$

Where c is the constant defined as:

$$\{ \Sigma^{\neq}_j \gamma_{0j}^2 \lambda j^{-1} \}^{-1} \{ \Sigma (Y_i - \bar{Y})^2 \}^{1/2}$$

and where $\Sigma^{\neq}$ denotes the summation over only those values, of I whose vectors have been retained in the procedure above. Suppose for example, we retain all the vectors, the first element of $b^{\neq}$ is infant the Least Squares coefficient.
According Webster et al. ([W+74]), the residual sum of squares for the modified Least Squares equation can be written as

Residual sum of square (SSE) :
$= \{ \Sigma (Y_i - \bar{Y})^2 \} \{ \Sigma^{\neq}_j \gamma_{0j}^2 \lambda j^{-1} \}^{-1}$
$= - C \{ \Sigma (Y_i - \bar{Y})^2 \}^{1/2}$

Apply this to the Least Squares solution as a check, we would find out that the result is true. Again, if we perform the parallel equation when the vector corresponding to the smallest Latent Root dropped, we will observe that there are considerable changes in the coefficient of the unbiased Least Squares values. However, the residual sum of squares (SSE) will usually increase a bit from its minimum value (attained by OLS). Thus the modified least equation can be represented in the sense of at least, as being nearly as good as the Least Squares equation.
A backward elimination procedure suggested by Webster et al. ([W+74]), especially formula 6.1 can now be adopted. The residual sum of squares that results from deletion of unwanted data X_l, $l = 1, 2. \ldots\ldots r$ from the modified Least Squares equation can be evaluated as:

$$\{ \Sigma (Y_i - \bar{Y})^2 \} \{ t_{oo} - t^2_{io} / t_{ij} \}^{-1}$$

Where, $t_{pq} = \Sigma^{\neq}_j (\gamma_{Pj} \lambda_{qj} / \gamma_j)$
And again where $\Sigma^{\neq}$ denotes the summation over entries retained in the modified least square calculation. According to Webster et al. ([W+74]), "The key advantage of this method is that, by unmasking the effect of the non – predictive near singularities, the true influences of the independent variables on the dependent variable are more clearly represented."

2.3 Ridge Regression Analysis

Among the numerous methods developed for obtaining biased estimator is the Ridge Regression which is obtained by solving a slightly modified version of the normal equation.
To calculate the Ridge estimate of the parameter in the model:

$$y_i = \beta_o + \beta_i X_{i1} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

we first consider the standardize regression model:

$$y_i' = \beta_i' X_i + \ldots + \beta_p' X'_{ip} + \varepsilon_i ,$$

This model does not employ an intercept term; this is because the least square calculations would always yield an estimated intercept of zero if any intercept parameter were used. The new parameters β_1', β_2', $\ldots$ β_p' are related to those in the original model β_0, β_1, $\beta_2, \ldots$ β_p. Also

$$\beta_j = (S_y / S_{xj}) \beta'_j , j = 1, 2, 3, \ldots\ldots p$$
and
$$y'_i = 1/(\sqrt{n} - 1) [(y_i - y) / S_y]$$
$$x'_{ij} = 1/(\sqrt{n} - 1) [(x_{ij} - x_j) / S_y]$$

In matrix form

$$y = \begin{pmatrix} y'_1 \\ y'_2 \\ . \\ y_n \end{pmatrix} \quad X = \begin{pmatrix} x'_{11} & x'_{12} & . & . & . & x'_{1p} \\ x'_{21} & x'_{22} & . & . & . & x'_{2p} \\ . & . & . & . & . & \\ x'_{n1} & x_{n2} & . & . & . & x'_{np} \end{pmatrix}$$

$$X'X = \begin{pmatrix} 1 & r_{x1x2} & r_{x1x3} & . & . & . & r_{x1xp} \\ r_{x2x1} & 1 & r_{x2x3} & . & . & . & r_{x2xp} \\ . & . & . & . & . & . & \\ r_{xpx1} & r_{xpx2} & . & . & . & . & 1 \end{pmatrix}$$

$$X'Y = \begin{pmatrix} r_{yx1} \\ r_{yx2} \\ r_{yxp} \end{pmatrix}$$

The ridge point estimate of the parameter β_1', β_2', $\ldots$ β_p' of the standardized regression model is denoted by:

$$b'_{j.R} = \begin{pmatrix} \beta'_{1.R} \\ \beta'_{2.R} \\ . \\ \beta'_{p.R} \end{pmatrix} = (X'X + kI)^{-1} X'Y \qquad (4)$$

Where k is a biasing constant $k \leq 0$, then the Ridge point estimate of parameter β_1', β_2', $\ldots$ β_p' in the original model are:

$$b_{j.R} = [(S_y / S_x)] b'_{j.R} \qquad j = 1, 2, \ldots p$$
$$b_{0.R} = \bar{y} - b_{1.R} \bar{x}1 - b_{2.R} \bar{x}2 - \ldots - b_{P.R} \bar{x}P$$

A special note on k (the biasing constant) is that: when it is equal to zero the Ridge point estimates become the Least Squares point estimates. It thus follows from (4) that,

$E(b_{j.R}) = E[\,(X'X + kI)^{-1} X'Y\,] = (X'X + kI)^{-1} X\, E(Y)$

Recall that $E(Y) = X\beta$, hence

$E(b_{j.R}) = Var\,(X'X + kI)^{-1} X'Y$
$\sigma^2 (X'X + kI)^{-1} X'X (X'X + kI)^{-1}$ (covariance matrix of $\boldsymbol{b}^{\neq}$)

The Ridge estimator is thus biased though it can be shown that the variance of the elements of $b_{j.R}$ are less than those of the least square estimator. According to Björkstrom ([Bjö01]), Ridge Regressors is known to have favorable properties as shown by Hoerl and Kennard ([HK70]). $R\,\hat{\boldsymbol{\beta}}$ has smaller mean square error (MSE) than the ordinary least squares estimator which provided k small enough and the standard regression model holds. Marquardt ([Mar70]) also pointed out that the Ridge regressors are known as shrinkage estimator.

The main problem centers on the selection of numerical value of the arbitrary k (the biasing constant). Recall that the Least Squares estimation procedure is unbiased i.e. $\mu_0 = \beta_j$. *If* $k > 0$ the Ridge Regression procedure is not unbiased i.e $\mu_{bj.R} \neq \beta_j$. If we define the bias of the Ridge estimation procedure to *be* $\mu_{bj.R} - \beta_j$. Employing the mean square error (MSE), we can compare a biased and an unbiased procedure. We then define the bias of the ridge estimation procedure to *be* $\mu_{bj.R} - \beta$. The MSE of an estimation procedure is the average square deviation of different possible point estimates from the unknown parameter. This is equal to the sum of squared bias of the procedure and its variance.

If we denote the bias estimator of β by $\boldsymbol{b}^{\neq}$ say: which has a smaller variance than the unbiased estimators, b, the total MSE of the Ridge estimator $b^{\neq}$ is defined as:

$MSE(b^{\neq}) = E(b^{\neq} - \beta)^2 = \text{var}(b^{\neq}) + [E(b^{\neq}) - \beta]^{-2}$
$= var(b^{\neq}) + (biased\ in\ b^{\neq})^2$
$\sigma^2\ Tr[(X'X + kI)^{-1} X'X (X'X + kI)^{-1}] + k^2 \beta' (X'X + kI)^{-2} \beta$
$\sigma^2\ \Sigma \boldsymbol{\lambda_j} / (\boldsymbol{\lambda_j} + k)^2 + k^2 \beta' (X'X + kI)^{-2} \beta$

Where $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, ... $\boldsymbol{\lambda}_p$ are the eigenvalues of X ' X

The variance here is the average of the square deviation of the different possible point estimates from the mean of all possible point estimates. If the procedure is unbiased, the mean of all estimates is the parameter being estimated. That is, when the bias is zero, the MSE and the variance of procedure are the same.

Therefore, the MSE of the unbiased Least Squares estimation procedure for testing βj is the variance σ^2_{bj} var (b) while that of Ridge estimation procedure is

$var(b) + (bias\ in\ b)^2 = \sigma^2_{bj.R} + (\mu_{bj})^2$

In general the Ridge point estimates are less affected by multicollinearity than the Least Squares.

3. RESULTS AND DISCUSSION

Each method discussed in this paper was tested using data on measures of money supply. The performances of Principal Component Regression, Latent Root Regression and Ridge Regression on the data set are summarized in table below. This evaluation was done using MSE values, where the method that produced the smallest value was the best method, performance of all the methods are determined using an efficiency test to find out which method was the best in specified cases. The *R* - Square values show the models' capability to fit the present data using those methods.

Table 1: The Comparative Statistics

	PCA		LRR		RR	
	X_3	X_4	X_3	X_4	X_3	X_4
R-Square	0.8952	0.9692	0.8023	0.9704	0.9814	0.9798
Adj. R^2	0.8690	0.9615	0.8526	0.9614	0.9707	0.9680
MSE	2.0188E+011	3.0383E+011	2.8828E+011	2.9156E+011	58374892539	3.25958E11
RMSE	4.4931E+005	5.5121E+005	4.3391E+005	5.3997E+005	241609	570927

With the use of Principal Component Regression analysis only six variables were selected out of the eleven variables, whereas by the use of the Latent Root Regression seven variables were selected which include those six selected by the Principal Component approach. Using M1 (X_3) as the dependent variable it was found that 86.9% of the total variation in the dependent variable was explained by the fitted model. The use of the Latent Root regression with X_3 as the dependent variable shows the adjusted R- square as 0.852552 implying that 85.3% of the total variation in the dependent variable is explained by the fitted model. While the use of Ridge Regression with X_3 as dependent variable six variables were selected by applying SAS (Statistical Analysis System**)** at the same time, all of the independent variables together explain 98.14% of the dependent variables can be seen through the

R- Square in table. One interesting thing is that the first six variables selected by this approach were also selected by Principal Component, this confirms the fact that the two approaches are possibly alike and that those six variables are actually good measures of money supply. With the R - Square value (0.9814) in the Ridge Regression which is larger than the R - Square value in the Principal Component 0.8952, is good evidence that the performance of the six variables selected by the Ridge Regression in measuring the variation in the dependent variable shows an improvement on the Principal Component approach.

Considering the values obtained in table which both has M2(X_4) as the dependent variable. The earlier table shows the adjusted R- Square as 0.9615 which implies that 96.2% of the total variation in M2(X_4) is explained by the stated model. This is on the consideration of the first six variables as the independent variable. While in the latter the adjusted R - Squared is 0.9614 implying that 96.14% of the total variation in the dependent variable M2(X_4) is explained by the fitted model. Looking at these results, this further confirm that the addition of the variable have not contributed positively to the variation in the dependent variable. With the decrease in the Mean Square Error (MSE), the earlier fact has been confirmed.

This result is quite significant in that, for the first analysis, the comparative statistics show adjusted R - Square from the Principal Component Regression approach as higher than that of the Latent Root Regression approach. Ridge regression performed best compared to the other two methods with increase in the adjusted R - Square of 0.9707, the second analysis shows the same thing for M2 (X_4). In the three cases, it is found that there is decrease in the MSE of the Ridge Regression approach as compared to the other two approaches which signifies the efficiency of the a particular approach – Ridge Regression analysis.

This implies that, if all the three approaches are properly made use of , the Ridge Regression more efficient to the other two methods which gives MSE = 58374892539 for M1(X_3)and MSE = 3.25958e11for M2 (X_4) followed by Principal Component with MSE = 2.0188E+011 for M1(X_3) and MSE = 3.0383E+011for M2 (X_4) and Latent Root with MSE = 2.8828E+011 for M1(X_3)and MSE = 2.9156E+011 for M2 (X_4) respectively. Ridge Regression method is considered the best since it shows the lowest MSE values for both M1(X_3) and M2(X_4). This MSE value indicates to what extent the slope and intercept are correctly estimated. So, the goal is to obtain an MSE value close to zero.

5. CONCLUSION

This paper has critically examined the three different methods of handling multicollinearity in economic data. The work has primarily established that, the Mean Square Error (MSE) of the Ridge Regression as the least of the three methods adopted, is the best compared with the others.

However, the observed adjusted R – Square for the approaches were found to be high i.e. 0.9680, 0.9615 and 0, 9614 in order of RR, PCA and LR which implies that the models were well fitted.

It is therefore recommended that the approach be used for removal of multicollinearity among independent variables.

For further exercises, it is recommended for any future research wishing to conduct a study in this area to try the use of the Partial Least Squares Regression (PLSR) analysis, which is another bias estimation technique to see if there exists any difference in the approach.

REFERENCES

[AA09] **Y. M. Al-Hassan, M. M. Al-Kassab** - *A Monte Carlo Comparison between Ridge and Principal Components Regression Methods*. Applied Mathematical Sciences, Vol. 3, 2085 - 2098, 2009.

[Bjö01] **A. Björkström** - *Ridge Regression and Inverse Problem. Research Report*. Sweden Stockholm University. 2001.

[BS99] **A. Björkström, R. Sundberg** - *A Generalized View on Continuum Regression*. Scandinavian Journal Statistics**. 25** : 17 –30, 1999.

[ER11] **M. El-Dereny, N. I. Rashwan** - *Solving Multicollinearity problem using Ridge Regression Models*. Int.J. Contemp. Math. Sciences, vol 6, no.12 p585 600, 2011.

[GM77] **R. F. Gunst, R. L. Mason** - *Biased estimation in regression: An evaluation using MSE*. Journal of the American Statistical Association, 72, 616-628, 1977.

[GM80] **R. F. Gunst, R. L. Mason** - *Regression Analysis and its Application: A Data-oriented Approach*. New York: Marcel Dekker, 1980.

[GP09] **D. N. Gujarati, D. C. Porter** - *Basic Econometrics*. Fourth edition. McGraw-Hill, 2009.

[HK70] **A. E. Hoerl, R. W. Kennard** - *Ridge Regression: Biased Estimation to Nonorthogonal Problems*. Technimetrics. 12: 56 –67,1970.

[Joh03] **J. Johnston** - *Econometrics Methods*. New York.4th ed. McGraw-Hill Int. 2003.

[Mar70] **D.W. Marquardt** - *Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation*. Technometrics, 15, 591-612,1970.

[Mor76] **D. F. Morrison** - *Multivariate Statistics Methods*. New York: McGraw-Hills lnt. 1976.

[MP90] **D. C. Montgomery, E. A. Peck** - *Introduction to Linear Regression Analysis*. 2nd. New York: John Wiley & Son. 1990.

[MS75] **D. W. Marquardt, R. D. Snee** - *Ridge regression in practice.* American Statistician Vol. 29 (1) P. 3–20, 1975.

[N+07] **A. Norliza et. al** - *Comparing Three Methods Of Handling Multicollinearity Using Simulation Approach.* Matematika, Vol. 22, Number 2, 2007.

[The71] **H. Theil** - *Principle of Econometrics*. New York: John Wiley & Sons. 1971.

[WBK80] **R. E. Welsch, D. A. Belsley, E. Kuh** - *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley and Sons. 1980.

[W+74] **J. T. Webster, R. F. Gunst, R. L. Mason** - *Latent root Regression analysis*. Technometrics Vol.16 P. 513-522, 1974.

[W+76] **J. T. Webster, R. F. Gunst, R. L. Mason** - *A comparison of least squares and latent root regression estimators*. Technometrics, 18: 75–83, 1976.