MAYFEB Journal of Mathematics - ISSN 2371-6193
Vol 3 (2017) - Pages 45-52

A Comparison Analysis of Shrinkage Regression Methods of Handling Multicollinearity Problems Based on Lognormal and Exponential Distributions

Umar Usman, Department of Mathematics, Usmanu Danfodiyo University, Sokoto. Nigeria, uusman07@gmail.com

Yahaya Zakari, Department of Mathematics, Usmanu Danfodiyo University, Sokoto. Nigeria, zakariyahaya007@gmail.com

Shamsuddeen Suleman, Department of Mathematics, Usmanu Danfodiyo University, Sokoto. Nigeria, suleman.shamsuddeen@udusok.edu.ng

Faruk Manu, Department of Mathematics, Usmanu Danfodiyo University, Sokoto. Nigeria, farukebbi@yahoo.com

***Abstract*-Handling multicollinearity problem in regression analysis is very important because the existence of multicollinearity among the predictor variables inflates the variances, and confidence interval of the parameter estimates which may lead to lack of statistical significance of individual independent variables, even though the overall model may have significance difference. It is also mislead p-values of the parameter estimate. In this paper, several regression techniques were used for prediction in the presence of multicollinearity which include: Ridge Regression (RR), Partial Least Squares Regression (PLSR) and Principal Component Regression (PCR). Therefore, we investigated the performance of these methods with the simulated data that follows lognormal and exponential distributions. Hence, Mean square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) were obtained. And the result shows that PLSR and RR methods are generally effective in handling multicollinearity problems at both lognormal and exponential distributions.**

Keywords: Multicollinearity, ridge regression, partial least square regression, principal component regression.

I. INTRODUCTION

In several linear regression and prediction problems, the independent variables may be many and highly collinear. This process is called multicollinearity and it is known that in this case the ordinary least squares (OLS) estimator for the regression coefficients or predictor based on these estimates may give very poor results [1].

The problem of multicollinearity in regression analysis can have effects on least squares estimated regression coefficients, computational accuracy, estimated standard deviation of least squares estimated regression coefficients, t-test, extra sum of squares, fitted values and predictions, and coefficients of partial determination. These problems can be remedied using some method of estimation or some modifications of the method of least squares for estimating the regression coefficients. Thus, the problem of multicollinearity can occur in both simple linear regression and multiple linear regressions [2]. According to [3], there are various procedures for dealing with

multicollinearity, some of these include Principal Component Regression (PCR), Partial Least Square Regression (PLSR), Ridge Regression (RR), e.t.c. Three regularized regression methods were compared by Root Mean Square Error and Root Mean Square Error Cross Validation on real data. The data consists of the Gross Domestic Product Per Capita (GDPPC) in Turkey. Ordinary Least Squares regression (OLS) and Ridge Regression (RR) were found to be best because the value of RMSE is minimum while partial least squares regression (PLSR) is best because RMSECV is minimum. Finally, they concluded in their study that Partial Least Squares Regression (PLSR) was the superior model in terms of the prediction ability as compared to the other regularized models [4]. Also, prediction methods of the RR, the PCR and the PLSR using Monte Carlo simulation study were compared by [3], the authors used predictors 2,4,6, and 50 with sample size 20,30, 40, 60, 80 and 100. Thus, they concluded that the RR is found to be the best for low number of regression (when sample size lie between 20 to 100), PCR is the best when number of observations are greater than number of regressors in the model and PLSR perform better results as compared to the other two prediction methods when have a number of regressors.

The [5] presented two techniques: PCA and PLSR for dimension reduction purpose when regressors are highly correlated. The PCA technique is used without the consideration of the correlation while the PLSR technique is applied based on the correlation using simulated data. They concluded that PLSR technique is more effective to the PCA technique for dimension reduction purpose. However, [6] compared the PLSR, RR and PCR as an alternative procedure for handling multicollinearity problem. The authors performed a Monte Carlo simulation to evaluate the effectiveness of these three procedures. Also, Mean Squared Errors (MSE) was calculated. Their results showed that the RR is more efficient when the number of regressors is small, while the PLSR is more efficient than the others when the number of regressors is moderate or high. Biased regression method PCR, PLSR and RR which stabilize the variance of the parameter estimate to overcome the problems of multicollinearity were compared by [7]. They used different levels of correlations to simulated data that follows normal and uniform distributions to estimate the regression coefficients by PCR, PLSR and RR methods. However, they compared the three methods by using symmetric loss functions such as, mean square errors (MSE), root mean square errors (RMSE), mean absolute errors (MAE) and mean absolute percentage errors (MAPE). Based on their study, they observed that PLSR has a lower measure of accuracy when data follows normal distribution while RR shows better results in uniform distribution. They finally recommended that these methods can be applied to the same distributions used in their study by varying the sample sizes and equally be used to look at the behaviours of other distributions other than those used in their research work.

II. METHODOLOGY

In this section three prediction methods: Partial Least Squares Regression (PLSR), Principal component Regression (PCR) and ridge Regressions are described briefly which are used in our study to remove the problem of multicollinearity.

1. Partial Least Square Regression

The PLSR searches for a set of components (called latent vectors) that performs a simultaneous decomposition of X and Y with the constraints that this components explain as much as possible the covariance between $X$and Y. In this method, the component was extracted from which the rest of the components are extracted in such a way that they are uncorrelated (orthogonal). How this algorithm functions will now be described to show how the PLS method works. The first is defined as:

$$t_i = W_{11}X_1 + W_{12}X_2 + \cdots + W_{ip} = \sum W_{ij}\, X_j \tag{1}$$

Where, X_j are the explanatory variables, Y is the dependent variables.

The W_{ij} is the coefficient:

$$W_{ij} = \frac{cov(X_j, Y)}{\sqrt{\sum_j^p cov(X_j, Y)^2}}, \text{ j=1, 2, 3…}p \tag{2}$$

From which it can be deduced that in order to obtain W_{ij} the scalar product (X_j, Y) must be calculated for each $J = 1, 2 \ldots P$.

Calculating the second component is justified when the single component model is inadequate i.e. when the explanatory power of regression is small and another component is necessary. The second component is denoted by t_2 and it will be a linear combination of the regression residues of X_j variables on components t_1 instead of the original variables. In this way, component orthogonality is assured. To do this, the residual for the single component regression was calculated according to the equations below:

$e_1 = Y - \hat{Y} = Y - \beta_1 t_1$with

$$\beta_1 = \frac{cov(y_i, t_i)}{\|t_1\|^2} \tag{3}$$

The second component is obtained as:

$$t_2 = W_{21}e_{11} + W_{22}e_{12} + \cdots + W_{2p}e_{1p} \tag{4}$$

With $W_{2j} = \frac{cov(e_{ij}, e_1)}{\sqrt{\sum_j^p cov^2(e_{ij}, e_1)}}$, j=1, 2, 3…$p$ (5)

The residuals e_{ij} are calculated by computing the simple regression of xjon$t1$,

$X^*_j = \alpha_j t_j,\ j = 1, 2, \ldots, p$therefore,

$$e_{ij} = X_j - X^*_j = X_j - \alpha_j t_j \tag{6}$$

Where, the estimators of the regression coefficients have been calculated thus:

$$\alpha_j = \frac{cov(x_j, t_1)}{\|t_1\|^2} \tag{7}$$

Now with e_i and e_{ij}, only the scalar products have to be computed $cov(e_i, e_{ij})$,

for j = 1 . . .P, to be able to compute t_2.

To construct subsequent components, the same steps was performed as for the two previous components. This iterative procedure is continued until the number of components to be retained is significant.

2. Principal Components Regression

The Principal Component Regression is a biased estimation technique in handling multicollinearity. It performs least squares estimation on a set of new variables called the Principal Components of the correlation matrix. The results in estimation and prediction are superior to ordinary least squares (OLS).

Suppose, the regression equation may be written in matrix form as

$$Y = XB + e \tag{8}$$

where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated and e represents the errors or residuals.

2.1 PC Regression Basics

In ordinary least squares, the regression coefficients are estimated using the formula

$$\hat{B} = (X'X)^{-1}X'Y \tag{9}$$

Note that 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 transformed to their principal components. Mathematically written according to the equation below:

$$X'X = PDP' = Z'Z \tag{10}$$

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

Thus, the new variables Z has been created as weighted averages of the original variable X. This is nothing new since we are used to using transformations such as the logarithm and the square root on the data values prior to performing the regression calculations. Since these new variables are PC, their correlations with each other are all zero. If variables X_1, X_2, and X_3 are used the result will be Z_1, Z_2, and Z_3.

Severe multicollinearity will be detected as very small Eigenvalues. To get rid the data of the multicollinearity, the components (the z's) associated with small eigenvalues will be omitted. Usually, only one or two relatively small eigenvalues will be obtained. For example, if only one small eigen value were detected on a problem with three independent variables, we would omit Z_3 (the third principal component).

When regress Y on Z_1 and Z_2, multicollinearity is no longer a problem. Then the result was transformed back to the X scale to obtain estimates of B. These estimates were biased, but the size of these biases was compensated by the decrease in variance. That is, the mean squared error of these estimates is less than that for least squares.

$$\hat{A} = (Z'Z)^{-1}Z'Y = D^{-1}Z'Y \tag{11}$$

Because of the special nature of principal components. Notice that this is ordinary least squares regression applied to a different set of independent variables.

3. Ridge Regression

When multicollinearity exists, the matrix $X'X$, where X consists of the original regressors, becomes nearly singular. Since $Var(\beta) = \delta^2(X'X),^{-1}$ and the diagonal elements of $(X'X)^{-1}$ become quite large, this makes the variance of β to be large. This leads to an unstable estimate of β when OLS is used.

3.1 Steps in Performing Ridge Regression

STEP I**:**

Consider the following regression model:

$$Y_i = \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon \tag{12}$$

Where, $\beta_1, \beta_2, \beta_3$ etc. are the parameters of the model and ε are random terms.

STEP II:

Standardize data by subtracting each x observation from its corresponding mean and dividing by its standard deviation i.e. $\frac{x_i - \mu_i}{\sqrt{\delta_i}}$

STEP III:

Arrange the predictors into convenient matrix. Suppose we have n observations of k predictors, this will be a $n \times k$matrix x. And arrange the key parameters into a β. So that viewing the response variable as an n-vector, our model becomes:

$$y = x\beta + \varepsilon \tag{13}$$

Where, ε is now a vector of the random noise in the observed data vector$Y$.

Note: the least square parameter β_{LS} can be estimated by finding the parameter values which minimized the sum square residuals i.e.

$SSR = \sum(Y - X\beta)'(Y - X\beta)$. The solution turns out to be a matrix equation,

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{14}$$

Where, X' is the transpose of the matrix$X$.

According to [8], the potential instability in using the least squares estimator could be improved by adding a small constant λ to the diagonal entries of the $X'X$ marix before taking its inverse.

The result is the Ridge regression estimator

$$\hat{\beta} = (X'X + \lambda I)^{-1}X'Y \tag{15}$$

Where I is the $p \times p$ identity matrix and $X'X$ is the correlation matrix of independent variables values of lambda lie in the range (0 and 1). When $\lambda = 0$, $\hat{\beta}_{RIDGE}$ becomes $\hat{\beta}_{OLS}$.Obviously, a key aspect of ridge regression is determining what the best value of the constant that is added to the main diagonal of the matrix $X'X$should be to maximize prediction. There are many procedures for determining the best value. The simplest way is to plot the values of each $\hat{\beta}_{RIDGE}$versus λ. The smallest value for which each ridge trace plot shows stability in the coefficient is adopted [9].

4. Comparative Study of PLSR, PCR and RR on Simulated Data

In this section we introduce the several measures of a model's fit to the data and of predictive power used in this paper.

i. Mean Square Error (MSE)

The MSE measures the average of the squares of the errors or deviations i.e the difference between the estimator and what is estimated.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (16)$$

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

ii. Root Mean Square Error (RMSE)

The RMSE is a measure of how well the model fits the data. It is defined as:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(yi-\hat{y}i)^2}{n}} \quad (17)$$

where the $\hat{y}_i$ are the values of the predicted variable when all samples are include in the model formation, and n is the number of observations.

iii. Mean Absolute Error (MAE)

The MAE is a quantity used to measure how close predictions are to the eventual outcomes.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (18)$$

It is an average of the absolute errors. i.e. $|e_i| = |f_i - y_i|$, where f_i is the prediction and y_i is the true value.

iv. Mean Absolute Percentage Error (MAPE)

The MAPE is a measure of prediction accuracy of a forecasting method in statistics. It usually expresses accuracy as a percentage, and is defined by the formula:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \quad (19)$$

Where, A_t is the actual value and F_t is the forecast value.

The difference between A_t and F_t is divided by the Actual value A_t again. The absolute value in this calculation is summed for every forecasted point in time and divided by the number of fitted points n. Multiplying it by 100 makes it a percentage error.

SIMULATION SETTINGS

We compared the PLSR, PCR and RR prediction methods on simulated data that follows lognormal and exponential distributions with parameters mean (μ) = 0 and δ^2 = 1. The numbers of variables are 5 with 250 observations. Thus, in the simulation, Monte Carlo study was performed by considering different levels of multicollinearity [10], [11, 12] and [7].

Also, a dependent variable is generated by using the equation.

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon, n = 1,2, \dots, p \quad (22)$$

where $\varepsilon_i$is a normal pseudo random number with mean zero and variance σ^2. The simulation was also replicated 160 times in order to obtain the approximate distribution considered in the study in real life situation. The correlations between the variables are $\rho_{12} = 0.2, \rho_{13} = 0.5, \rho_{14} = 0.3, \rho_{15} = 0.3, \rho_{23} = 0.5, \rho_{24} = 0.9, \rho_{25} = 0.7, \rho_{34} = 0.5, \rho_{35} = 0.7, \rho_{45} = 0.9$.

Variance Inflation Factor (VIF) was also used to check the presence of multicollinearity in the data simulated. Therefore the results of simulations are listed below:

TABLE 1: Mean Square Error

Regressions	Probability distributions	
	Lognormal	Exponential
PCR	5.103552	1.163262
PLSR	**3.513678**	**1.15274**
RR	5.195772	1.166902

It has been revealed from Table1 that partial least squares regression (PLSR) has high predictive ability at both lognormal and exponential distributions which make it better than both PCR and RR.

TABLE 2: Root Mean Square Error

Regressions	Probability distributions	
	Lognormal	Exponential
PCR	2.28204	1.078546
PLSR	**1.874481**	**1.062869**
RR	2.279423	1.080233

From the results of Table 1, it has been observed that partial least squares regression has high predictive abilities at both lognormal and exponential distributions, which means that PLSR performed better than both ridge and principal component regressions.

TABLE 3: Mean Absolute Error

Regressions	Probability distributions	
	Lognormal	Exponential
PCR	1.330704	**0.4449592**
PLSR	1.326204	0.8349363
RR	**1.167424**	0.798196

Table 3, revealed that ridge regression has minimum value than both PLSR and PCR when data follows lognormal distribution while when data follows exponential distribution principal component regression has predictive ability than both PLSR and RR. Therefore, Ridge Regression performed better at lognormal distribution but at exponential distribution principal component performed better.

TABLE 4: Mean Absolute Percentage Error

Regressions	Probability distributions	
	Lognormal	Exponential
PCR	2.076961	5.669157
PLSR	2.285785	5.752208
RR	**2.074079**	**5.035345**

From Table 4, which shows that Ridge Regression the least value at the both lognormal and exponential distributions. i.e. Ridge regression performed better at both distributions than PLSR and RR.

CONCLUSION

In this paper, PLSR, PCR and RR have been applied to the simulated data and it shows that PLSR and RR methods are generally effective in handling multicollinearity problems both lognormal and exponential distributions.

ACKNOWLEDGMENT

We appreciate all who have contributed in one way or the other to the success in the pursuit of this research work.

REFERENCES

[1] R.F. Gunst and R.L. Mason, "Some considerations in the evaluation of alternate prediction equations, *Technometrics,* 1979. **21**, 55-63.

[2] T.H. Wannacott and R.J. Wannacott, "*Regression a Second Course in Statistics,"* John Wiley and sons, USA*,* 1981.

[3] N.J. Adnan, M.H. Ahmad and R. Adnan, (2006). A Comparative Study on some Methods for Handling Multicollinearity Problems. *Matematik.* **22**(2): 109-119.

[4] O. Yeniay and A. Goktas, (2002). A Comparison of Partial Least Squares Regression with other Prediction Methods. *Hacettepe Journal of Mathematics and Statistics*. **31:** 99-111.

[5] S. Maitra and J. Yan (2008). Principal Component Analysis and Partial Least Squares: Two Dimension Reduction Techniques for Regression. *Casualty Actuarial Society Discussion paper Program.* 79-90.

[6] M. El-Fallah and A. El-Salam, (2014). A Note on Partial Least Squares Regression for Multicollinearity (A Comparative Study). *International Journal of Applied Science and Technology*. **4**(1):163-169.

[7] U. Usman, Y. Zakari and Y. Musa, (2017). A Comparative Study on the Prediction Performance Methods of Handling Multicollineariy Based on Normal and Uniform Distributions. *Journal of Basic and Applied Research International.* **21**(1). In Press.

[8] A.E. Hoerl and R.W. Kennard (1970). Ridge Regression Applications to Non-orthogonal Problems,*Technometrics*.**12**(1)**:** 69-82.

[9] R.H. Mayers, (1990) *Classical and Modern Regression with Applications.*2nd edition, Duxbury press.

[10] G.C. Mc Donald and A. Galarneau, (1975). A Monte Carlo Evaluation of some Ridge-Type Estimators. Journal of American Statistical Association, 70, 407-416.

[11] S. Arumairajan and P. Wijekoon (2014). Improvement of Ridge Estimator when Stochastic Restrictions are Available in the Linear Regression Model. *Journal of Statistical and Econometric Methods,* **3**. 35-48.

[12] S. Arumairajan and P. Wijekoon (2015). Optimal Generalized Biased Estimator in Linear Regression Model. *Open Journal of Statistics*, 5. 403-411.