PHYSICAL SCIENCES

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

REGULARIZATION TECHNIQUES IN MULTIPLE LINEAR REGRESSION IN THE PESENCE OF MULTICOLLINEARITY

[1]Oyegoke, O. A., [2]Oyeyemi, G. M., [3]Adeleke, M. O. and [4]Kolawole, R. O.

[1]Department of Statistics, Osun State Polytechnic, Iree, Nigeria
[2,3]Department of Statistics, University of Ilorin, Ilorin, Nigeria
[4]Department of Building Technology, Federal Polytechnic Offa, Nigeria

Corresponding E-mail: gmoyeyemi@gmail.com

Manuscript Received: 04/04/2019 *Accepted:11/02/2020* *Published: March 2020*

ABSTRACT
Multicollinearity has been a serious problem in regression analysis. Ordinary least square (OLS) regression may result in high variability in the estimates of the regression coefficients in the presence of multicollinearity. Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression (RR), and Partial Least Squares (PLS) methods are well established methods that reduce the variability of the estimates by shrinking the coefficients and at the same time produce interpretable models by shrinking some coefficients. The performances of LASSO, Ridge Regression, PLS and OLS estimators were evaluated using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in the presence of multicolinearity using Monte Carlo Simulation. The simulations were done for different sample sizes: n (10, 50, 100, 150) and levels of multicollinearity: Mild (0.1 – 0.3), Low (0.4 – 0.6) and High (0.7 - 0.9). OLS had poor parameters estimate and produced wrong inferences, LASSO estimator is the best, while PLS is most efficient when the number of variable is greater than sample size.

Keywords: *Multicollinearity, Least Absolute Shrinkage and Selection Operator, Ridge Regression, Partial Least Squares and Estimators.*

INTRODUCTION

Multiple linear regression is a widely used statistical technique that allows us to estimate models that describe the distribution of a response variable with the help of a number of other variables usually called explanatory variables, or independent variables. The use of multiple regression mainly regards the interpretation of the regression coefficients. In case of independent coefficients, the least-squares solution gives stable estimates and useful results. However, data are not always "well behaved". We often come across cases where the explanatory variables are nearly collinear. This condition is called multicollinearity and is one of the most often encountered problems in econometrics. The major problem with multicollinearity is that it leads to estimates with inflated variances in the estimation of regression coefficients and thus unacceptably large prediction intervals (Zou and Hastie, 2005).

High estimate of variances (and therefore high estimated standard errors) also mean small observed test statistics. This implies that the analyst will accept too many null hypotheses. Estimates of standard errors and parameters tend to be sensitive to changes in the data and the specification of the model (Oyeyemi *et al.,* 2015). In addition, the least-squares estimates are usually inflated with wrong signs- though they remain the best linear unbiased estimates (BLUE). Note that, if the aim of the analyst is to generate forecasts, and if it is assumed that the multicollinearity problem will not be different for the forecast period, then multicollinearity may be considered not to be a problem at all. This is because multicollinearity will not affect the forecasts of a model but only the estimation of the coefficients (Knight and Fu, 2000, Muhammad *et al.,* 2013).

In order to detect the presence of collinear variables, many diagnostics have been proposed in the literature, for instance the condition number, variance inflation factors, variance decomposition proportions etc. Approaches to remedy the problem of multicollinearity have also been proposed. Model specification, variable selection, and biased estimation are some of them. In 1970 Arthur Hoerl and Robert Kennard published a paper on ridge regression, also known as the biased estimation method, that became the most commonly used method for the remedy of multicollinearity. Hoerl and Kennard's method was in fact a crude form of regularization, a technique developed by Andre Tikhonov (Tikhonov and Arsenin, 1977). Ridge regression involves the introduction of some bias into the regression equation in order to reduce the variance of the estimators of the parameters.

The bias is introduced by the use of a ridge constant which controls the extent to which ridge estimates differ from the least squares estimates. Depending on the method used to calculate the optimum ridge constant, i.e. the constant that provides the greatest amount of explained variance in the parameter estimators, different ridge estimators are defined. Ridge estimators have been proposed by many authors. McDonald and Galarneau (1975) proposed an estimator whose squared length equals an estimated squared length of β. Based on the mean square error property of the ridge estimator Hoerl, Kennard and Baldwin (1975) and others have also proposed ridge estimators. Furthermore, considering the Bayesian approach, Lindley and Smith (1972), Lawless and Wang (1976) and others have also introduced ridge estimators.

The classical linear regression model is

$$y_i = X_i\beta + \epsilon_i \qquad \text{.....(1)}$$

Where y_i and ϵ_i are (column) vectors of length n, X is a n × p (where p= k+1) matrix, and β is a vector of length p. The vector β of coe□cients is estimated so as to minimize the errors ϵ_i . If the number of data points n exceeds the number of predictors p, it is generally possible to find a β that gives a perfect fit (that would be y_i=X_i β, with no error, for all data points i = 1, . . . , n), and the usual estimation goal is to choose the estimate $\hat{\beta}$ that minimizes the sum of the squares of the residuals $e_i = y_i - X_i\hat{\beta}$ The sum of squared residuals is

$$SSE = \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2 \;;$$

the $\hat{\beta}$ that minimizes it is called the least squares estimate and can be written in matrix notation as

$$\hat{\beta}\ ols = (X^T X)^{-1}(X^T Y \qquad \text{.....(2)}$$

The errors ϵ_i come from the normal distribution with mean 0 and variance σ^2 . This variance can be estimated from the residuals, as;

$$\widehat{\sigma^2} = \frac{1}{n-k} SSE = \frac{1}{n-k} \sum_{i=1}^{n} (y_i - X_i\hat{\beta})^2 \qquad \text{.....(3)}$$

with n^k rather than n^1 in the denominator to adjust for the estimation of the k-dimensional parameter β.

If there is no linear relationship between the regressors, they are said to be orthogonal. Multicollinearity is a case of multiple regression in which the predictor variables are themselves highly correlated. If the goal is to understand how the various X variables impact Y, then multicollinearity is a big problem. Multicollinearity is a matter of degree, not a matter of presence or absence. Under severe multicollinearity the ordinary least squares estimators are imprecise and unstable. There are several methods available in literature for detection of multicollinearity. By observing correlation matrix, variance inflation factor (VIF) or Eigenvalues of the correlation matrix, one can detect the presence of multicollinearity.

The Ordinary least square method tends to produce estimates that are imprecise and unstable, leading to poor prediction. In order to prevent this

difficulties of ordinary least squares method (OLS), Hoerl and Kennard (1970) suggested the ridge regression as an alternative procedure to the least square method in regression analysis. The ridge regression which is the linear transformation of the least square method is based on adding a biasing constant k to the diagonal of X'X matrix before computing β's. Therefore, ridge regression is given by

$$\beta_{ridge} = (X'X+ kI)\text{-}1X'Y. \quad K\geq 0 \quad .....(4)$$

where k is the ridge parameter and I is the identity matrix. The value of k is appropriate between the interval of (0, 1). Note that if k =0 the ridge estimator becomes the ordinary least square.The values of k will be selected by the analyst. The corresponding values of the ridge parameter produces different regression coefficient. As the value of k increases from zero the smaller the variance, the greater the biased introduced. It is always difficult to select the optimal value of k that produces the stable regression coefficients. Some various methods are used in selecting the appropriate value of k.

The main step in the ridge regression analysis is to select a value of K and to obtain the corresponding estimates of the regression coefficients. There are several methods of selecting the appropriate value of K. The procedure can be summarized as follows:

i. Ridge trace: Hoerl and Kennard (1970) suggested a graphical method called ridge trace to select the value of the ridge parameter K. This is a plot of the values of the regression coefficient plotted against the range of the values of K.

ii. Fixed point: Hoerl, Kennard and Baldwind (1975) suggest estimating K by

$$KHKB= \frac{p\sigma^2}{\sum_{i=1}\beta^2_{(0)}} \quad .....(5)$$

With p as the number of predictors and β(0) is the least square estimate of the regression parameters.

iii. Iterative method: Hoerl and Kennard (1976) proposed the following iterative procedure for selecting K: start with the initial estimate of K in (1) denote this value by K0,then calculate

$$K1 = \frac{p\sigma^2}{\sum_{i=1}\beta^2_{(k0)}} \quad .....(6)$$

Then use K1 to calculate

$$K2 = \frac{p\sigma^2}{\sum_{i=1}\beta^2_{(k1)}} \quad .....(7)$$

Repeat this process until the difference between two successive estimates of K is negligible, in this work, an interval of 0.0001 between successive ridge parameter was used. The one which gives the minimum MSE of regression parameter is selected as the best model.

MATERIALS AND METHODS

This study used Monte Carlo study design to simulate the data. This includes data generation, estimation and methods of computing various criteria for assessing the performance of estimators. The simulation and analysis was carried out using R statistical software (www.cran.org).

The ridge regression procedure (Hoerl and Kennard, 1970) is based on the matrix X'X+ kI, I denoting the identity matrix and k being a positive scalar parameter. It is a procedure that can be used in "ill-condition" situations where correlations between the various predictors in the model cause the X'X matrix to be close to singular. In particular, we can obtain a point estimate with a smaller mean square error.

Hoerl and Kennard (1970) suggested that in order to control inflation and general instability associated with the least squares estimates, one can use

$$\hat{\boldsymbol{\beta}}(k)= (X'X+ kI)^{-1}X'Y \quad k\geq 0 \quad ..... (8)$$

Note that the LS estimator is a member of this family with k = 0, the ridge estimator, though biased, has lower mean square error than the BLUE (best linear unbiased estimator). Unfortunately, this mean-squared error is a function of the unknown parameters that we are trying to estimate. Thus, an estimator with low MSE will be close to the true parameter. One property of the least squares estimator β^ that is frequently noted in the ridge regression literature is (Tibshirani, 1996)

$$E(\beta'\beta) = \beta'\beta + \sigma^2 tr(X'X)^{-1} > \beta'\beta + \frac{\sigma^2}{\lambda_k} \quad .....(9)$$

where λ_k is the minimum eigenvalue of X'X. Thus, if the data are collinear, and λ_k is small, this implies that the expected squared length of the least squares coefficient vector is greater than the squared length of the true coefficient vector. In addition, the smaller the λ_k, the greater the difference.

The mean of the ridge estimator is given by

$$E(\hat{\boldsymbol{\beta}}(k)) = G_k X' E(Y) \quad .....(10)$$

Where $G_k=(X'X+ kI)^{-1}$

$$E(\hat{\boldsymbol{\beta}}(k)) = G_k X' X\boldsymbol{\beta} \quad .....(11)$$

$$= Z_k \hat{\boldsymbol{\beta}}$$

Note that when k=0 then Z_k=I hence E(β (k))=β, but when k≠0, β ̂(k) provides a biased estimate of β.

The bias of the estimator β ̂(k) is given by Bias(β ̂(k))=-kG_k β. Indeed, we know that the bias of an estimator b* is defined as

Bias(b*)=E(b*)-β

Consequently, the bias of $\hat{\boldsymbol{\beta}}$(k) is

Bias($\hat{\boldsymbol{\beta}}$(k))=E($\hat{\boldsymbol{\beta}}$(k))-β=Z_k $\hat{\boldsymbol{\beta}}$-β

$=[(X'X+ kI)^{(-1)} X'X-I]\beta$

$=(X'X+ kI)^{-1} [X'X-(X'X+ kI)]\beta$

$=-kG_k \beta$

The variance-covariance matrix for the ridge regression estimators is

$$cov(\hat{\beta}(k)) = cov(Z_k\hat{\beta}) = Z_k cov(\hat{\beta}) Z_k$$

$$= \sigma^2 Z_k (X'X)^{-1} Z_k'$$

LASSO-type estimators are the techniques that are often suggested to handle the problem of multicollinearity in regression model. More often than none, Bayesian simulation with secondary data has been used. When the ordinary least squares are adopted there is tendency to have poor inferences, but with LASSO-type estimators which have recently been adopted may still come with its shortcoming by shrinking important parameters (James et al, 2009). Considering the classical Gaussian regression model (CGRM) given by (1).

With n observations and p features $X = [x_1 ,...,x_p]$, response y, coefficient vector β and (stochastic) error ε independently identically distributed Gaussian. In the recent time, most applications of statistics e.g health, finance etc. are often faced with p >> n, the true β vector is sparse, that's existence of many irrelevant predictors with coefficient equal to zero. Usually, the focus is to identify and select the important predictors and also obtain efficient estimates of their coefficients. However, the process of identifying and selecting subset of important features from thousands of features is generally difficult. As a remedial measure to this unavoidable problem, shrinkage or convex relaxation techniques like Least Absolute Shrinkage and Selection Operator (LASSO) is adopted.

Multivariate regression methods like Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) enjoy large popularity in a wide range of fields, including the natural sciences. The main reason is that they have been designed to confront the situation where there are possibly correlated predictor variables and relatively few samples (Pages and Tenanhaus, 2001, Xun and Liangjun, 2013). Comparative Molecular Field Analysis (ComFA) uses PLSR. Other applications range from statistical process control to tumour classification to spatial analysis in brain images to marketing. In the usual multiple linear regression (MLR) context, the least-squares solution for

$$Y= X\beta + \varepsilon$$

is given by $B=(X'X)^{-1} X'Y$

The problem often is that X'X is singular, either because the number of variables (columns) in X exceeds the number of objects (rows), or because of collinearities. Both PCR and PLSR circumvent this by decomposing X into orthogonal scores T and loadings P

$$X = T P$$

and regressing Y not on X itself but on the first a columns of the scores T. In PCR, the scores are given by the left singular vectors of X, multiplied with the corresponding singular values, and the loadings are the right singular vectors of X.

Akaike's Information Criterion (AIC) is based on maximum likelihood and penalty for each parameter. The general form is:

$$AIC = -2\log L + 2p$$

Where L is the likelihood and p is the number of parameters. In multiple regressions, this becomes:

$$AIC = n\log\left(\frac{RSS}{n}\right) + 2p + c$$

Where RSS is the residual sum of squares and c is constant.The best model by this criterion minimizes AIC.

Schwartz's Bayesian Information Criterion (BIC) is similar to AIC but penalizes additional parameters more. The general form is:

$$BIC = -2\log L + (\log n)p$$

Where n is the number of observation, L is likelihood and p is the number of parameters. In multiple regressions, this becomes:

$$BIC = n\log\left(\frac{RSS}{n}\right) + (\log n)p + c$$

Where RSS is the residual sum of squares and c is constant.The best model by this criterion minimizes BIC

RESULTS AND DISCUSSION

Simulation study was carried out to examining the finite sample performance for OLS, Ridge Regression, LASSO and Partial Least Square using both AIC and BIC as performance criteria. A number of independent variables were generated with different levels of multicollinearity: small (r = 0.1 – 0.3), mild (r = 0.4 – 0.6) and serious (r = 0.7 – 0.9) with different sample sizes n: (n = 10, 50, 100 and 150) from normal distribution. The simulation was repeated 1000 times for consistency. The number of independent variables considered in this study is fifteen (15), hence in one occasion the number of parameters to be estimated is more than the sample size (p > n). The four methods of estimation in multiple linear regression were used in the estimation of the model parameters and both the AIC and BIC of the fitted models are presented in Table 1 while Table 2 presents the overall performance of the four methods.

Table 1: Average AIC and BIC of the fitted model using the four methods

		Estimators							
		OLS		Ridge Regression		LASSO		PLS	
n	r	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC
10	0.1-0.3	NA	NA	79.499	83.130	94.951	98.582	**54.136**	**57.766**
	0.4-0.6	NA	NA	82.780	86.411	92.197	95.828	**50.049**	**53.680**
	0.7-0.9	NA	NA	85.783	89.414	63.810	67.441	**51.447**	**55.078**
50	0.1-0.3	158.927	181.871	161.018	183.963	**155.490**	**178.434**	167.744	190.688
	0.4-0.6	155.879	178.823	155.808	178.752	**146.924**	**169.868**	161.704	184.648
	0.7-0.9	156.594	179.538	154.615	177.559	**148.146**	**171.090**	159.213	182.157
100	0.1-0.3	294.757	326.019	295.390	326.652	**291.016**	**322.278**	305.768	337.030
	0.4-0.6	303.947	355.209	303.354	334.616	**297.552**	**328.814**	317.082	348.344
	0.7-0.9	302.743	334.006	299.514	330.816	**292.121**	**323.383**	304.334	335.796
150	0.1 0.3	438.716	474.844	438.915	475.042	**434.372**	**470.510**	451.614	487.771
	0.4-0.6	441.455	477.502	438.346	474.474	**434.512**	**470.640**	453.378	489.506
	0.7-0.9	446.754	482.882	440.663	476.811	**436.482**	**472.610**	450.132	486.260

Table 2: SUMMARY OF RESULTS BASED ON ESTIMATORS

R	Samples size (n)	Best
Low	10	PLS
	50	LASSO
	100	LASSO
	150	LASSO
Mild	10	PLS
	50	LASSO
	100	LASSO
	150	LASSO
High	10	PLS
	50	LASSO
	100	LASSO
	150	LASSO

Table 1 shows both the AIC and BIC of the fitted models using the four methods. It is of interest to note that both criteria agreed in selecting the best method in all the cases considered. It can be observed that PLS performed better at all the three levels of multicollinearity with sample size of n = 10; and LASSO performed better at all the three levels of multicollinearity with sample sizes of: n = 50, 100 and 150. Therefore, LASSO appears to have best overall performances among all the methods considered as shown in Table 2.

CONCLUSION

The OLS, Ridge regression, LASSO and Partial least square (PLS) estimators were compared in the presence of multicollinearity in linear regression model. OLS broke down when number parameters are more than sample size (p > n) and had poor parameters estimate in the presence of multicollinearity even when p < n hence produces wrong inferences. LASSO estimator has the best performance followed by ridge regression while the partial least square (PLS) is only preferred method when the sample size is less than number of parameters (p > n).

REFERENCES

Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12, 55–67*.

Hoerl, A. E., Kennard, R. W. and Baldwin, K. F. (1975). Ridge Regression: Some Simulations. *Communication in Statistics-Theory and Methods, 4 (2), 212 – 220.*

Hoerl, A. E. & Kennard, R. W. (1976). Ridge regression: Applications to nonorthogonal problems. *Technometrics, 12, 69– 82*.

James, G. M., Radchenko, P. and Lv, J. (2009). Connections between the dantzing selector and Lasso, *Journal of the Royal Statistical Society, Series B.71,121- 142.*

Knight, K. and Fu, W. (2000). Asymptotic for Lasso-type estimators. *Annals of Statistics 28, 1356- 1378*.

Lawless, J. F. and Wang, P. (1976). A Simulation Study of Ridge Regression and other Regression Estimators. *Communication in Statistics-Theory and Methods, 5 (4), 112 – 122.*

Lindley, D. V. and Smith, A. F. M. (1972). Bayes Estimates for the Linear Model. *Journal of the Statistical Society Series B (Methodological), 34 (1), 422 – 433.*

McDonald, G. C. and Galarneau, D. I. (1975). A Monte Carlo Evaluation of some Ridge-Type Estimators. *Journal of the American Statistical Association, 70 (350), 407 – 416.*

Muhammad, I., J. Maria and A. R. Muhammad, 2013. Comparison of shrinkable Regression for Remedy of Multicollinearity Problem. *Middle – East Journal of Scientific Research 14(4): 570 – 579.*

Oyeyemi, G.M., Ogunjobi, E.O. and Folorunsho, A. I. (2015). Performance of Shrinkage Methods. *International Journal of Statistics and Applications, 5(2): 72-76.*

Pages, J. and Tenenhaus, M. (2001). Multiple factor analysis combined with PLS path modeling.Application to the analysis of relationships between physicochemical variables, sensory profiles and hedonic judgments. *Chemometrics and Intelligent Laboratory Systems, 58, 261–273.*

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B. 58, 267-288.*

Tikhonov, A. N. and Arsenin, V. Y. (1977). Solutions of Ill-Posed Problems. Wiley, New York.

Xun L. and Liangjun, S. (2015). Shrinkage Estimation of Dynamic Panel Data Models with Interactive Fixed Effects. School of Economics, Singapore Management University. 1 - 45

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B 67, 301 – 320.*