International Journal of Engineering & Technology, 7 (4.30) (2018) 36-38

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Comparison of OLS and Ridge Regression Methods in the Presence of Multicollinearity Problem in the Data

N S M Shariff[1]*, H M B Duzan[2]

[1,2] Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM),
Bandar Baru Nilai 71800, Nilai, Negeri Sembilan, Malaysia
**Corresponding author E-mail: nurulsima@usim.edu.my*

Abstract

The presence of multicollinearity will significantly lead to inconsistent parameter estimates in regression modeling. The common procedure in regression analysis that is Ordinarily Least Squares (OLS) is not robust to multicollinearity problem and will result in inaccurate model. To solve this problem, a number of methods are developed in the literatures and the most common method is ridge regression. Although there are many studies propose variety method to overcome multicolinearity problem in regression analysis, this study proposes the simplest model of ridge regression which is based on linear combinations of the coefficient of the least squares regression of independent variables to determine the value of k (ridge estimator in ridge regression model). The performance of the proposed method is investigated and compared to OLS and some recent existing methods. Thus, simulation studies based on Monte Carlo simulation study are considered. The result of this study is able to produce similar findings as in existing method and outperform OLS in the existence of multicollinearity in the regression modeling.

Keywords: *Multicollinearity; OLS; Ridge Regression.*

1. Introduction

Ordinary Least Squares (OLS) is the Best Linear Unbiased Estimator (BLUE) in investigating the relationship between explanatory and response variables in the regression modeling. The OLS is only applicable when all regression assumptions are satisfied and some of them are; errors in the model are distributed with normal distribution with zero mean and a constant variance and no high correlation problem among the explanatory (independent) variables. Multicollinearity is defined as conditions on which some or all explanatory variables have large influence on others explanatory variables. This critical issue might happen if the analysis contains large sets of data with several numbers of explanatory variables and this will affect to the existence of multicollinearity problem. In the presence of multicollinearity, the regression assumptions are invalid and as such, the OLS cannot be preceded in the next stage of estimation. Otherwise, the results of parameter estimates and inference under OLS procedure will be insignificant and unreliable.

Due to such problem, there are quite number methods of estimations to overcome multicollinearity problem in regression analysis. [1, 2] are the first to introduce ridge regression method by adding small positive quantities (denoted by letter k in many studies (see [3]) to the diagonal of the matrix $\mathbf{X}^T\mathbf{X}$ where $\mathbf{X}^T\mathbf{X}$ is matrix of explanatory variables) and it is shown can minimize the biased estimates and mean squared error (MSE) of the model. [3] reviews some existing estimators from years 1964 to 2014 with a wide variety of techniques to estimate k and make a comparison with OLS model in the paper. It is found that most of the models outperform OLS and concludes that the generalized ridge regression is the best model based on the smallest MSE value.

Due to interest in choosing the most appropriate k in ridge regression method, this study proposes another technique with k can be formed as a linear combination of coefficients of determination of explanatory variables (see [4]). The performance of the proposed method is investigated and comparison is made to OLS and some existing methods by using Variance Inflation Factor (VIF) and MSE criterion.

2. Methodology

Let us assume that the matrix of response variable $\mathbf{Y}$ $(n´\ 1)$ can be predicted as a linear relationship with explanatory (independent) variables $\mathbf{X}$ $(n´\ p)$ as follows

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \tag{1}$$

where $\mathbf{b}$ represents p coefficients for $\mathbf{X}$ and $\mathbf{e}$ are residuals with the assumptions of $\mathbf{e}$ are $E(\mathbf{e}) = 0$ and $Var(\mathbf{e}) = s^2\mathbf{I}_n$. In OLS procedure, the parameter estimates of $\mathbf{b}$ in (1) is then given by

$$\hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{2}$$

2.1. Ridge regression model

The OLS estimate in (2) is invalid due to large deviation in $\mathbf{b}$. The $\hat{\mathbf{b}}$ is said to be unbiased but inconsistent. To overcome this problem, the positive value of k is added to the diagonal elements in $\mathbf{X}^T\mathbf{X}$ matrix in (2) (see [1, 2]) to minimize the impact of high correlation in explanatory variables [5]. Hence, the ridge regression model is

$$\hat{\boldsymbol{\beta}}_R = \left(\mathbf{X}^T\mathbf{X} + k\mathbf{I}_n\right)^{-1}\mathbf{X}^T\mathbf{Y} \qquad (3)$$

According to [1], the value of the ridge estimator *k* in (3) will provide a smaller value of MSE compared to OLS estimator.

2.2. Different types of ridge estimator

Several recent methods of ridge estimators are considered to estimate the value of *k* (See [6-8])

$$\hat{k}_2 = \frac{1}{p}\sum_{j=1}^{p}\left(\frac{\lambda_j\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \lambda_j\hat{\beta}_j^2}\right), \quad j = 1,2,\dots,p \qquad (4)$$

$$\hat{k}_4 = median\left(\frac{\lambda_j\hat{\sigma}^2}{(n-p)\hat{\sigma}^2 + \lambda_j\hat{\beta}_j}\right), \quad j = 1,2,\dots,p \qquad (5)$$

[9] Then proposes the following estimator

$$\hat{k}_{12} = \left(\prod_{j=1}^{p}\frac{\lambda_j\hat{\sigma}^2}{\left(n-p\right)\hat{\sigma}^2 + \lambda_j\hat{\beta}_j^2}\right)^{\frac{1}{p}}, \quad j = 1,2,\dots,p \qquad (6)$$

where λ_j is eigenvalue of the $\mathbf{X}^T\mathbf{X}$ matrix, $\hat{\sigma}^2$ is the estimates of standard error σ^2, and $\hat{\beta}_j$ is the estimates of j^{th} element in $\hat{\mathbf{b}}$ as defined in (2).

2.3. Proposed ridge regression estimator

An alternative method of the existing ridge method is proposed where *k* is estimated by using coefficient of determination of explanatory variables in the regression model [4]. From (2), the scaled variables $\mathbf{X}$ are assumed to be in the form of correlation matrix. To illustrate this, let (1) is written as

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p + e \qquad (7)$$

Then, consider

$$\bar{Y} = \beta_0 + \beta_1\bar{X}_1 + \beta_2\bar{X}_2 + \cdots + \beta_p\bar{X}_p \qquad (8)$$

by subtracting (7) to (8), it will produce

$$Y - \bar{Y} = \beta_1\left(X_1 - \bar{X}_1\right) + \beta_2\left(X_2 - \bar{X}_2\right) + \cdots + \beta_p\left(X_p - \bar{X}_p\right) + e \qquad (9)$$

And note that

$$S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ and } S_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)^2;\ j = 1,2,\dots p.$$

The parameterized model with transformed variables is then given by

$$Y_i^* = \beta_1^*X_{i1}^* + \beta_2^*X_{i2}^* + \dots + \beta_p^*X_{ip}^* + e_i^* \qquad (10)$$

With the standardized variables Y_i^* and X_{ij}^* are simplified as the following form

$$Y_i^* = \frac{1}{\sqrt{n-1}}\frac{(Y_i - \bar{Y})}{S_y}, \quad X_{ij}^* = \frac{1}{\sqrt{n-1}}\frac{(X_{ij} - \bar{X})}{S_j};\ j = 1,2,\dots p \qquad (11)$$

where $\beta_j^* = \frac{\beta_jS_j}{S_y};\ j = 1,2,\dots,p$.

Then, the estimates of β is given by

$$\hat{\mathbf{b}}^* = \left(\mathbf{X}^{T}\mathbf{X}^*\right)^{-1}\mathbf{X}^{T}\mathbf{Y}^* \qquad (12)$$

Note that, the values in the matrix $\mathbf{X}^{T}\mathbf{X}^*$ can be written as

$$\sum_{i=1}^{n} x_{ij}^{*2} = \sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{S_j\sqrt{n-1}}\right)^2 = \frac{\frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}{n-1}}{S_j^2} = 1\ ;\quad j = 1,2,\dots.p \text{ ; and}$$

$$\sum_{i=1}^{n} x_{ij}^{*\prime}x_{ik}^* = \sum_{i=1}^{n}\left(\frac{x_{ij} - \bar{x}_j}{S_j\sqrt{n-1}}\right)\left(\frac{x_{ik} - \bar{x}_k}{S_j\sqrt{n-1}}\right)$$

$$= \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2\sum_{i=1}^{n}(x_{ij} - \bar{x}_k)^2}} = r_{ij},\ \ j;k = 1,2,\dots p;\ j \neq k \qquad (13)$$

Where r_{ij} is the correlation coefficient between X_i and X_j.
In the regression model with *p* number of coefficients, the diagonal elements of matrix $C = \left(\mathbf{X}^{T}\mathbf{X}^*\right)^{-1}$ is written as $C_{jj} = \left(1 - R_j^2\right)^{-1}, j = 1,2,\dots,p$ with R_j^2 is the coefficient of determination in the regression of an explanatory variable X_j. The parameter estimates under OLS procedure are invalid when there is a perfect linear relationship among explanatory variables in the model. As R_j^2 tends to 1 ($R_j^2 \to 1$), the *j*th diagonal element of $\left(\mathbf{X}^{T}\mathbf{X}^*\right)^{-1}$ will be very huge. Since $var(\hat{\beta}_j) = \hat{\sigma}^2C_{jj}$, then $R_j^2 \to 1$ as $var(\beta_j) \to \infty$. Due to such problem, the new estimates of β is proposed by adding *k* in (12):

$$\hat{\mathbf{b}}_R^* = \left(\mathbf{X}^{T}\mathbf{X}^* + k\mathbf{I}\right)^{-1}\mathbf{X}^{T}\mathbf{Y}^* \qquad (14)$$

With *k* is simulated and the appropriate estimates of *k* for 4 explanatory variables is (see [4]):

$$\hat{k} = 0.174R_1^2 + 0.170R_2^2 + 0.194R_3^2 + 0.199R_4^2. \qquad (15)$$

3. Monte Carlo simulation study

In this study, the performance of existing ridge estimators are compared with OLS and the proposed ridge model via Monte Carlo simulation study. The chosen methods are OLS; Ridge estimator with $\hat{k}_2$; Ridge estimator with $\hat{k}_4$; Ridge estimator with $\hat{k}_{12}$; and the proposed estimator of *k* as in (15).

Following the simulation procedure as in [10], the data generation process of this study is

$$X_{ij} = (1- g^2)^{\frac{1}{2}} z_{ij} + g z_{ij}, \quad i = 1,2,...,n, \;\; j = 1,2,...,p \tag{16}$$

where z_{ij} are independent and identically normal distributed, and g is defined such that correlation between any two independent variables. The chosen $g$are $g=0.7$, 0.8, and 0.9 (for low, moderate, and high correlations between the variables, respectively) and n = 50 with number of p is 4 and 2,000 number of simulations. This study considers p=4 due to the reason of minimizing multicollinearity effect on MSE since less number of independent variables are correlated when the variables is reduced.
The response y is determined by using the following model

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + ... + b_p x_{ip} + e_i, \quad i = 1,2,...,n \tag{17}$$

where e_i are independent and identically normal distributed, $N(0,1)$, and b_0 is assumed to be identically zero. The values of the parameters are set to be $\mathbf{b}^{\mathbf{T}}\mathbf{b} = 1$ indicating that if MSE is the function of b, s^2 and k, then MSE is minimized upon selection of this coefficient vector [11]. The measure of performance are investigated by using the values of VIF and MSE as

$$\text{VIF} = \left(1- R_j^2\right)^{-1} \text{and} \quad MSE\left(\hat{b}_j\right) = \text{var}(\hat{b}_j) + Bias\left(\hat{b}_j\right)^2 \tag{18}$$

The use of VIF measures is to indicate the severity of multicollinearity problem in the data. The common rule of thumb of the presence of multicollinearity is when VIF value is larger than 5. MSE measures the spread of the parameter estimates and the difference between the true values of the parameter. The smaller MSE illustrates the better result of the parameter estimates.

3.1. Results and discussion

This section presents the result of the Monte Carlo simulation study. To save some space in this paper, the important results that might be useful to summarize the findings are presented. At first, the correlation values are computed to illustrate the presence of high dependency among explanatory variables (when $g=0.9$) and it is reported in Table 1. The results indicate the presence of high correlated explanatory variables with all correlation values are larger than 0.8.

Table 1: Correlation Values of Explanatory Variables

Explanatory variables	X_1	X_2	X_3	X_4
X_1		0.8061	0.8069	0.8977
X_2	0.8061		0.8070	0.8976
X_3	0.8069	0.8070		0.8987
X_4	0.8977	0.8976	0.8987	

Table 2: The VIF, and MSE results when $g=0.9$

Model Type	VIF$_1$	VIF$_2$	VIF$_3$	VIF$_4$	AMSE
OLS	5.6725	5.6511	5.723	15.0284	0.6001
k_2	0.9802	0.9627	0.9747	0.9025	0.0402
k_4	0.9259	0.9823	0.9811	0.9836	0.0451
k_{12}	0.9255	0.9805	0.9884	0.9803	0.0270
The proposed model	0.9577	0.9838	0.9816	0.9844	0.0202

Note: VIF$_1$, VIF$_2$, VIF$_3$ and VIF$_4$ refer to the VIF values for each explanatory variable (p=4), respectively. AMSE is the average of MSE of all explanatory variables.

Table 2 presents the result of MSE and VIF of OLS, existing methods and proposed method for $g=0.9$. The VIFs of OLS are larger than 5 for all explanatory variables indicate the presence of severe dependency among explanatory variables in the data. The existing and proposed models yield reasonable values of VIF. For MSE results, the proposed method provides better result than other estimators with the smallest value of MSE indicating the results is satisfactorily in the regression modeling.

4. Conclusion

Another approach with different method of ridge estimator is considered with the aim of reducing the multicollinearity in the data. The performance of estimators is investigated using Monte Carlo simulation study with 2,000 numbers of replications using two types of measures, MSE and VIF. The finding of this study shows that the proposed ridge estimator performs better than other estimator, as multicollinearity exists in the data. It can be concluded that the proposed ridge estimator can produce reliable results as other type of estimators.

Acknowledgement

The authors acknowledge for the financial support from **(PPPI/FST/0217/051000/11118)** from Universiti Sains Islam Malaysia.

References

[1] Hoerl AE & Kennard RW (1970), Ridge regression: applications to nonorthogonal problem. *Technometrics,* 12(1), pp. 69-78.

[2] Hoerl AE & Kennard RW (1970), Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), pp. 55-67.

[3] Duzan H & Shariff NSM (2015), Ridge regression for solving the multicollinearity problem: review of methods and models. *Journal of Applied Sciences,* 15(3), pp. 392-404.

[4] Duzan H & Shariff NSM (2016), Solution to the multicollinearity problem by adding some constant to the diagonal. *Journal of Modern Appllied Statistical Methods,* 15(1)**,** pp. 752-773.

[5] Mansson K, Shukur G & Kibria BMG (2010), A simulation study of some ridge regression estimators under different distributional assumptions. *Communications in Statistics- Simulation and Computation,* 39(8), pp. 1639-1670.

[6] Kibria BMG (2003), Performance of some new ridge regression estimators. *Communications in Statistics- Simulation and Computation,* 32(2), pp. 419-435.

[7] Khalaf G & Shukur G (2005), Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods,* 34(5), pp. 1177-1182.

[8] Alkhamisi M & Shukur G (2008), Developing ridge parameters for SUR model. *Communications in Statistics - Theory and Methods,* 37(4), pp. 544-564.

[9] Muniz G, Kibria BMG & Shukur G (2012), On developing ridge regression parameters: a graphical investigation. *Department of Mathematics and Statistics* **1.**http://digitalcommons.fiu/math_fac/10

[10] Muniz G & Kibria BMG (2009), On some ridge regression estimators: an empirical comparisons. *Communications in Statistics- Simulation and Computation,* 38(3), pp. 621-630.

[11] Newhouse JP & Oman SD (1971), An evaluation of ridge estimators. *Rand Report,* No R-716-Pr, pp. 1-28.