International Journal of Engineering & Technology, 7 (4.30) (2018) 106-108

International Journal of Engineering & Technology

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

Research paper

An Application of Proposed Ridge Regression Methods to Real Data Problem

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 Ordinary Least Squares (OLS) is a common method to investigate the linear relationship among variable of interest. The presence of multicollinearity will produce unreliable result in the parameter estimates if OLS is applied to estimate the model. Due to such reason, this study aims to use the proposed ridge estimator as linear combinations of the coefficient of least squares regression of explanatory variables to the real application. The numerical example of stock market price and macroeconomic variables in Malaysia is employed using both methods with the aim of investigating the relationship of the variables in the presence of multicollinearity in the data set. The variables on interest are Consumer Price Index (CPI), Gross Domestic Product (GDP), Base Lending Rate (BLR) and Money Supply (M1). The obtained findings show that the proposed procedure is able to estimate the model and produce reliable result by reducing the effect of multicollinearity in the data set.

Keywords: *Macroeconomic Variables; Multicollinearity; OLS; Ridge Regression.*

1. Introduction

Multicollinearity refers to high strength correlation among the explanatory variables in regression analysis. This situation commonly occurs due to large number of explanatory variables are incorporated in the analysis. The well-known method of estimation that is Ordinary Least Squares (OLS) is unreliable because the violation of the independency assumption of explanatory variables. The high dependency in explanatory variables will cause huge value in the standard errors of parameter estimates and thus, OLS is no longer appropriate to be employed in modeling the data.

Several methods of estimation have been suggested in the literatures to overcome the multicollinearity problem. The most popular method is ridge regression that has been introduced by [1, 2]. The ridge regression is proposed by introducing a positive value of *k* to the diagonal of the matrix $\mathbf{X}^T\mathbf{X}$ (where $\mathbf{X}$ is matrix of explanatory variables) with the aim of minimizing the biased estimates and mean squared error (MSE) of the model. The *k* is also known as ridge estimator and there are a variety of methods to estimate *k* (see [3 - 5]). Some of them make comparison with OLS and find ridge estimators outperform OLS and conclude that the generalized ridge regression is the best model among all methods.

On the other hand, an alternative method of the existing ridge method is proposed in [6] where *k* is estimated by using coefficient of determination in the regression of an explanatory variable. This method has shown able to produce reliable results as in existing ridge methods in such away the dispersion of the standard error of the parameter estimates can be minimized. Due to such interest, the method will be applied to financial data. In the financial data, some explanatory variables maybe correlated due to high dependency among each other. This problem will have a vital influence on the results of statistical inference in the regression analysis. In preliminary analysis, the descriptive analysis and the correlation measures are analyzed to describe the data and illustrate the presence of multicollinearity. The next step is the modeling procedure to obtain the results of the parameter estimates and the inference of the model. The remaining section of this paper is organized as follows: Section 2 presents the data and model used in this study while Section 3 discusses the results followed by the conclusion is written in Section 4.

2. Data and model

The data of macroeconomic variables represented by interest rate (base lending rate (BLR)), inflation (consumer price index (CPI)), gross domestic product (GDP) and monetary supply (M1) and stock price are on interest in this study. The macroeconomic variables refer to explanatory variables which can be indicators to Malaysian’s economic that might affect the stock market movement. As such, the response (dependent variable) is stock price index (Kuala Lumpur Composite Index (KLCI)) with the length of the study is 15 years (2000-2015) and quarterly basis that yield 654 total number of observations.

2.1. Ridge regression model

Suppose that the data consists of *n* observations. Let Y_i be a scalar response and x_i be a vector of *p* explanatory variables. Then, a linear regression model is considered as follows

$$Y_i = x_i^T b + e_i \tag{1}$$

where β is a $p\times 1$ vector of parameters and e_i are scalar random errors. In matrix form, (1) can be written as

$$\mathbf{Y}=\mathbf{X}\beta+\mathbf{e} \tag{2}$$

In (1) and (2), the random errors are assumed to have zero mean and a constant variance; $E(\mathbf{e})=0$ and $Var(\mathbf{e})=\sigma^2\mathbf{I}_n$. As such, the OLS assumptions is fulfilled and thus the following estimates of β is obtained

$$\hat{\beta}=\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{3}$$

To minimize the impact of high dependency among **X** in (3), the value of k is added and the new equation is then given by

$$\hat{\beta}_R=\left(\mathbf{X}^T\mathbf{X}+k\mathbf{I}_n\right)^{-1}\mathbf{X}^T\mathbf{Y} \tag{4}$$

2.2. Proposed ridge regression estimator

This study uses the method in [6] to estimate *k* in (4). Here, *k* is estimated by using coefficient of determination in the regression of an explanatory variable. In this method, the scaled variables $\mathbf{X}$ in (3) are assumed to be in the form of correlation matrix. Consider (2) can be written as

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p+e \tag{5}$$

Taking average of *n* observations of (5)

$$\bar{Y}=\beta_0+\beta_1\bar{X}_1+\beta_2\bar{X}_2+\cdots+\beta_p\bar{X}_p \tag{6}$$

Then, consider the parameterized model with transformed variables is given by

$$Y_i^*=\beta_1^*X_{i1}^*+\beta_2^*X_{i2}^*+\ldots+\beta_p^*X_{ip}^*+e_i^* \tag{7}$$

with $Y_i^*=\dfrac{1}{\sqrt{n-1}}\dfrac{(Y_i-\bar{Y})}{S_y}$, $X_{ij}^*=\dfrac{1}{\sqrt{n-1}}\dfrac{(X_{ij}-\bar{X})}{S_j}$,

$S_y^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2$, $S_j^2=\dfrac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\bar{X}_j)^2$; and

$\beta_j^*=\dfrac{\beta_jS_j}{S_y};\ j=1,2,\ldots,p$

Then, the estimates of β is given by

$$\hat{\beta}^*=\left(\mathbf{X}^{T}\mathbf{X}^*\right)^{-1}\mathbf{X}^{T}\mathbf{Y}^* \tag{8}$$

Specifically, the values in the matrix $\mathbf{X}^{T}\mathbf{X}^*$ is written as

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

; and

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

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

where r_{ij} is the correlation coefficient between X_i and X_j.

In the *p*-variables case, the diagonal elements of $C=\left(\mathbf{X}^{T}\mathbf{X}^*\right)^{-1}$ is $C_{jj}=\left(1-R_j^2\right)^{-1}, j=1,2,\ldots,p$ with R_j^2 is the coefficient of determination in the regression of an explanatory variable on the remaining explanatory variables of the model. When there are high or perfect linear dependency among some or all independent variables in the model, the parameter estimates of OLS coefficients in (1) is unreliable and thus give inaccurate inference. Note that when 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 large. Since $var(\hat{\beta}_j)=\hat{\sigma}^2C_{jj}$, then $R_j^2\to 1$ as $var(\beta_j)\to\infty$. As such, the new estimate of β is proposed by adding *k* to overcome the multicollinearity problem:

$$\hat{\mathbf{b}}_R^*=\left(\mathbf{X}^{T}\mathbf{X}^*+k\mathbf{I}\right)^{-1}\mathbf{X}^{T}\mathbf{Y}^* \tag{9}$$

with *k* is simulated in [6] and the appropriate estimates of *k* for four explanatory variables is given by $\hat{k}=0.174R_1^2+0.170R_2^2+0.194R_3^2+0.199R_4^2$.

Others criterions of variables are investigated in this study are MSE and Variance inflation factor (VIF). The equation in (9) will minimize the value of MSE when the multicollinearity occurs. The VIF is the measure to indicate the presence of multicollinearity. The MSE and VIF are then computed as follows

$$\text{VIF}=\left(1-R_j^2\right)^{-1}\text{and}\quad MSE\left(\hat{\beta}_j\right)=\text{var}(\hat{\beta}_j)+Bias\left(\hat{\beta}_j\right)^2 \tag{10}$$

3. Results and Discussion

Table 1 illustrates the descriptive analysis of all variables used in this study. It can be seen that most variables provide similar results of measures of center that are mean and median values. It can be said that there is no peculiar observation that might affect these measures except for GDP and BLR. This result is supported by the slight larger values in standard deviation and kurtosis. The result shows GDP has negative skewness with fat tails distribution.

The presence of multicollinearity is investigated using correlation and it is presented in Table 2 and Figure 1. The high dependency among CPI, BLR and M1 can be seen from the results. The CPI and BLR are related to interest rates in economic point of view. The similar results are also observed for M1, BLR and CPI. Thus, the parameter estimation methods that encounter the mulicollinearity problem need to be employed to achieve the aim of the study.

Table 1: Descriptive analysis of variables

Variables	KLCI	CPI	GDP	BLR	M1
Mean	7.222	4.755	4.691	7.486	12.189
Median	7.039	4.545	5.300	6.310	12.051
Standard Deviation	0.715	0.513	2.818	3.143	0.873
Kurtosis	-0.147	0.563	3.220	0.360	-0.346
Skewness	1.067	1.544	-1.513	1.407	0.911

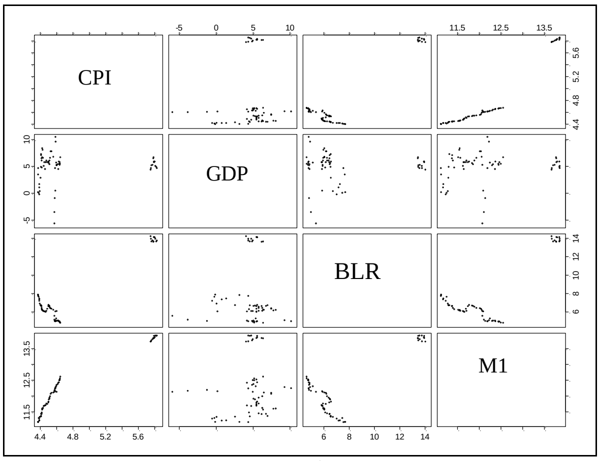

Fig. 1: Correlation plot of explanatory variable

Table 2: Correlation analysis of explanatory variables

Explanatory variables	CPI	GDP	BLR
GDP	0.111		
BLR	0.921*	0.041	
M1	0.952*	0.167	0.767

* Indicate the presence of multicollinearity with the value of correlation is higher than 0.9

Table 3: The results of estimation procedure

Estimator	Parameter	Coefficient	p-value	VIF	R^2	MSE
	b_1	-0.946	0.065	356.749*	0.979	0.107
OLS	b_2	0.015	0.008*	1.071		
	b_3	0.139	0.003*	79.086*		
	b_4	0.871	0.001*	130.816*		
Proposed ridge estimator	b_1	-0.864	0.009*	5.597*	0.925	0.079
	b_2	0.033	0.049*	0.169		
	b_3	0.560	0.001*	1.239		
	b_4	1.359	0.000*	2.052		

Note: b_1, b_2, b_3 and b_4 refer to the parameter estimates for CPI, GDP, BLR and M1 respectively.

* Indicate the parameter is significant at 5% level of significance.

* Indicate the presence of multicollinearity with the value of VIF is higher than 5

The result of estimation of OLS and the proposed method is shown in Table 3. Both methods provide different result and it can be seen that OLS provide significant results for all parameters indicating that all variables affect the stock market movement. The result of the proposed method however yields significant results for GDP, BLR, M1 but not for b_1 that represents CPI. This is due to the presence of high dependency between CPI and BLR, MI. In comparison to VIF, R^2 and MSE values, our proposed method provide smallest value whereby the sum squared of error is minimized by the value of *k* rather than in OLS and results in the value of MSE. The VIF values shown that the proposed method reduces the high dependency problem in the explanatory variables**.** Thus, the suggested method is able to estimate the model in the presence of high dependency of variables in the model.

4. Conclusion

This study proposes a ridge estimator as in [6] to solve the problem in regression analysis in the presence of high dependency among explanatory variables for the real data application. The proposed method is applied to investigate the relationship between macroeconomic variables and stock market movement. The estimation method of OLS and proposed method provides almost similar results and it is shown that the proposed method of estimation is able to produce consistent results as existing methods of estimation in the presence of multicollinearity in the data.

Acknowledgement

The authors are acknowledging 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), 69-78.

[2] Hoerl AE & Kennard RW (1970), Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 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), 392-404.

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

[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), 1639-1670.

[6] 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)**,** 752-773.