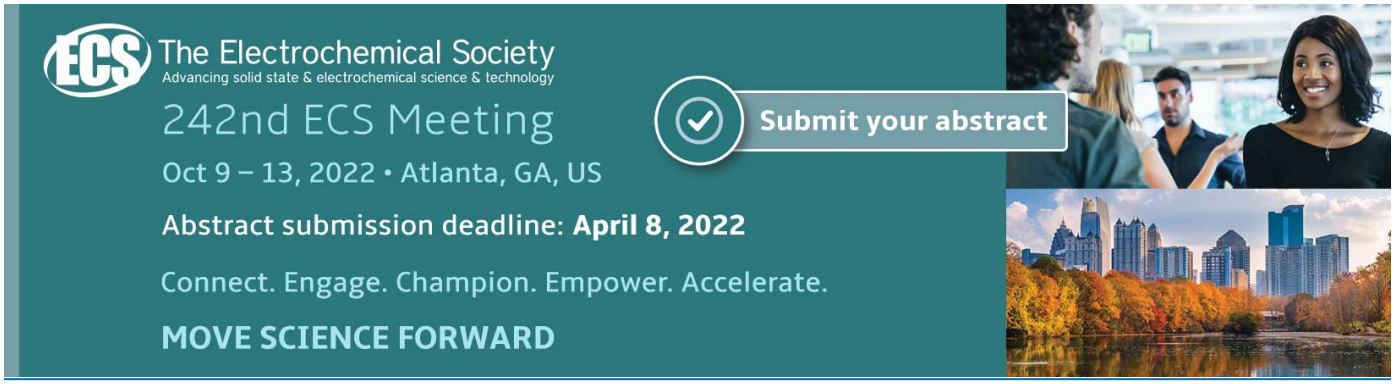
An application of robust ridge regression model in the presence of outliers to real data problem



An application of robust ridge regression model in the presence of outliers to real data problem

N S Md. Shariff and N A Ferdaos

Faculty of Science and Technology, Universiti Sains Islam Malaysia (USIM), Bandar Baru Nilai 71800, Nilai, Negeri Sembilan, Malaysia

E-mail: nurulsima@usim.edu.my, nuraqilahferdaos@yahoo.com

Abstract. Multicollinearity and outliers are often leads to inconsistent and unreliable parameter estimates in regression analysis. The well-known procedure that is robust to multicollinearity problem is the ridge regression method. This method however is believed are affected by the presence of outlier. The combination of GM-estimation and ridge parameter that is robust towards both problems is on interest in this study. As such, both techniques are employed to investigate the relationship between stock market price and macroeconomic variables in Malaysia due to curiosity of involving the multicollinearity and outlier problem in the data set. There are four macroeconomic factors selected for this study which are Consumer Price Index (CPI), Gross Domestic Product (GDP), Base Lending Rate (BLR) and Money Supply (M1). The results demonstrate that the proposed procedure is able to produce reliable results towards the presence of multicollinearity and outliers in the real data.

1. Introduction

Ordinary Least Squares (OLS) is the well-known technique that is used to find the linear relationship between variables in regression analysis. This OLS is said to be Best Linear Unbiased Estimator (BLUE) when all assumptions are satisfied: which are residuals in the model are identically and independently normally distributed with mean zero and a constant variance. These assumptions will be invalid in the presence of huge value in the standard errors of the estimated coefficients that is caused by multicollinearity and outliers in the data. Hence, OLS is inappropriate and results in inconsistent, inefficient and biased estimators of the model.

Multicollinearity is a situation where some of explanatory variables are highly correlated with others yielding inconsistent estimates. The analysis on huge data sets with large number of explanatory variables will lead to the presence of multicollinearity, especially in financial data. In the economic and financial data, some observed values maybe inconsistent from other observations in a dataset. These isolated or extreme values are termed as outliers and often have a large impact on the results of the statistical analyses in the regression model.

[1, 2] introduced ridge regression method to overcome multicollinearity problem in the data. However, this estimation process has a limitation where it is influenced by the presence of outliers. This problematic point will cause error rates inflated and slight distortion in parameter estimates. In many cases some data points will be deflected away from their expected values that are deemed reasonable. Therefore, an alternative procedure that is robust towards both problems is introduced in study with the illustration of the real data problem. The combination of GM-estimation technique that is adapted from [3] and ridge parameter is considered this study. Some descriptive statistics and the correlation analysis are computed in the initial stage of study followed by estimation process with aim of obtaining the parameter estimates with some inference results. The paper is then organized as follows: data and methodology are discussed in Section 2. A numerical example and results are presented in Section 3 and the conclusion is given in Section 4.

2. Data and methodology

The relationship between macroeconomic variables and stock price are on interest with the aim of illustration purposes of proposed method in the presence of outliers and multicollinearity in the data. The explanatory variables that represent macroeconomic variables are interest rate (base lending rate (BLR)), inflation (consumer price index (CPI)), gross domestic product (GDP) and monetary supply (M1). These variables are indicators or Malaysian's economic and it is believed to have relationship with stock market movement and thus, stock price index (Kuala Lumpur Composite Index (KLCI)) is used to be dependent variable in this study. The study period of variables is covered from 2000 until 2015 with quarterly basis which gives the total number of observation is 64.

2.1. Ridge regression method

Consider the following linear regression model:

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{e} \qquad (1)$$

where $\mathbf{Y}$ is a vector of n response (dependent) values, $\mathbf{X}$ represents $n \times p$ of explanatory (independent) variables with rank p, $\mathbf{b}$ is the vector of p coefficient for explanatory variables and $\mathbf{e}$ are random errors with the assumptions of zero mean and a constant variance, that are $E(\mathbf{e}) = 0$ and $Var(\mathbf{e}) = \sigma^2 \mathbf{I}_n$. Under OLS assumptions, the estimates of $\mathbf{b}$ in (1) is

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

In the presence of multicollinearity, the constant value of k is added to the diagonal elements in $\mathbf{X}^T\mathbf{X}$ matrix in equation (2) to reduce the dependency in explanatory variables and yield the ridge regression method with the following equation [1 , 2];

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

A combination of robust estimation methods and ridge parameter k is introduced in this study to solve both multicollinearity and outliers problem in the data.

2.2. Different Types of Estimator

There are many available literatures propose variety methods to find the value of *k*. See for example: [1 - 2, 4 - 10]. Based on the performance of [9], the technique of [8] is chosen to estimate *k*;

$$k_4 = \left(\prod_{i=1}^{p}\frac{1}{m_i}\right)^{\frac{1}{p}} \text{ with } m_i = \left(\frac{\hat{\sigma}^2}{\beta_i^2}\right)^{\frac{1}{2}} \qquad (4)$$

2.3. Proposed Robust Ridge Regression Estimator

Generalized M-estimator (GM) provides a good estimates in equation (1) where it filters both outliers in X and Y-directions by using the weight function $w(t) = \frac{\psi(t)}{t}$ where $\psi(t)$ is given by Huber function $\psi(t) = \begin{cases} t & ;|t| \le \varpi \\ \varpi\, sign(t) & ;otherwise \end{cases}$, where ϖ is set to 1.345 at 95% efficiency at normal distribution.

The GM estimator is given by $\hat{\mathbf{b}}_{Rob} = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$. By adopting the technique of [3], k^* is introduced in equation (3):

$$\hat{\mathbf{b}}_{RobR} = \left(\mathbf{X}^T\mathbf{X} + k^*\mathbf{I}_n\right)^{-1}\mathbf{X}^T\mathbf{X}\hat{\mathbf{b}}_{Rob} \qquad (5)$$

where $k^* = \dfrac{p.\hat{\sigma}_{Rob}}{\hat{\mathbf{b}}_{Rob}^T\hat{\mathbf{b}}_{Rob}}$, p and $\hat{\sigma}_{Rob}$ are the number of explanatory variables and robust scale, respectively. $\hat{\sigma}_{Rob}$ is the Median Absolute Deviation (MAD) and computed as $\hat{\sigma}_{Rob} = 1.4825\, median\left|\hat{\mathbf{e}} - median(\hat{\mathbf{e}})\right|$ and $\hat{\mathbf{e}}$ are estimated errors and obtained via $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\hat{\mathbf{b}}_{Rob}$.

3. Results and Discussions

Table 1 reveals the results of summary statistics for each variable in this study. All variables except GDP and BLR show closer value between mean and median as a measure of centres for mean and median indicating the absence of outliers' effects in the data. GDP and BLR provide a slight different value in both measures in the presence of outliers and it can be seen in figure 1. It is then strongly proven by large value of standard deviation and kurtosis especially for GDP. So, it is expected that GDP has fat tails and negatively skewed.

Before estimating the model, the presence of multicollinearity among the explanatory variables is investigated using correlation coefficients (see results in table 2 and figure 2). It can be seen that CPI has strong dependency with BLR and M1 but not with GDP. It well verse that CPI and BLR are linked because both are related to interest rates. Similar relationships are seen with M1, BLR and CPI. Thus, in view of the presence of the multicollinearity and outliers in the data, the parameter estimation procedure that relaxes the independence assumption should really be considered.

Table 1. Descriptive Statistics of variables.

Variables	KLCI	CPI	GDP	BLR	M1
Mean	7.2221	4.7546	4.6906	7.4856	12.1891
Median	7.0392	4.5449	5.3000	6.3100	12.0510
Standard Deviation	0.7148	0.5128	2.8184	3.1425	0.8733
Kurtosis	-0.1465	0.5638	3.2201	0.3604	-0.3463
Skewness	1.0668	1.5440	-1.5130	1.4068	0.9105

Table 2. Correlation Coefficient (*r*) of explanatory variables.

Explanatory variables	CPI	GDP	BLR
GDP	0.1108		
BLR	0.9212*	0.041	
M1	0.9523*	0.1666	0.7637

* indicate the presence of high multicollinearity due to large value of *r*.

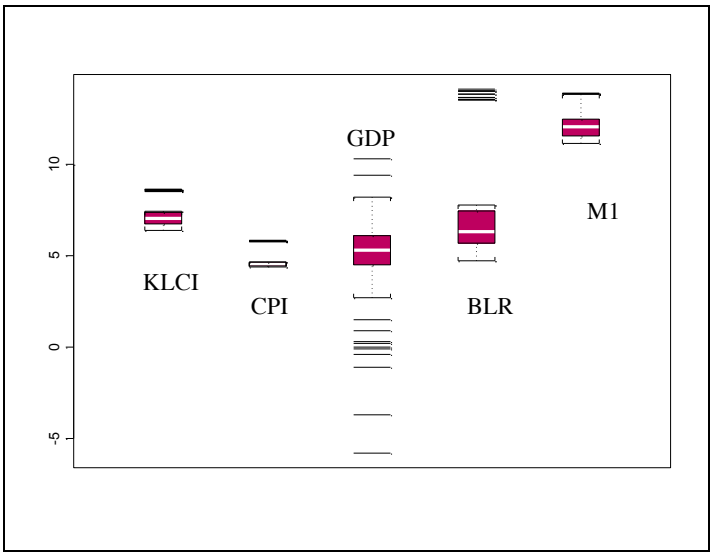


Figure 1. Boxplots of variables.

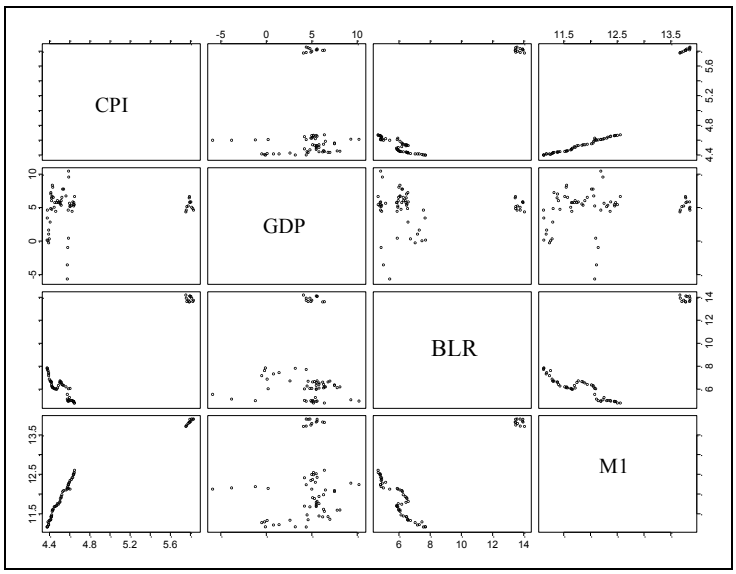


Figure 2. Correlation plot between explanatory variables.

The result of estimation is shown in table 3 and only CPI (refer to β_1 value) doesn't yield significant relationship towards KLCI for all estimation methods. This result however indicates that there is a strong relationship between GDP, BLR, M1 with stock market movement (KLCI). All estimates provide good R^2 and MSE values. In comparison with the R^2 and MSE values, our proposed method that is GM-estimator with $k = k^*$ provide smallest value and yet outperformed other estimation methods in dealing with multicollinearity and outliers in the data.

Table 3. The results from the estimation processes.

Estimator	Parameter	Coefficient	p-value	R^2	MSE
Ridge with $k = k_4$	β_1	-0.0699	1.0000	0.9632	0.0197
	β_2	0.0165	0.0000*		
	β_3	0.0712	0.0000*		
	β_4	0.5700	0.0000*		
GM-estimator with $k = k^*$	β_1	-0.0912	0.9226	0.9772	0.0092
	β_2	0.0143	0.0000*		
	β_3	0.0730	0.0000*		
	β_4	0.5786	0.0000*		
GM-estimator with $k = k_4$	β_1	-0.3201	0.6810	0.9636	0.0195
	β_2	0.0140	0.0000*		
	β_3	0.0907	0.0000*		
	β_4	0.6572	0.0000*		

Note: β_1 , β_2 , β_3 and β_4 refer to the estimated parameter for CPI, GDP, BLR and M1 respectively.
* indicate the parameter is significant at 5% level of significance.

4. Conclusion

This study suggest a combination of ridge regression and robust procedure to encounter multicollinearity and outliers in real data application. The proposed method is employed to study the relationship between stock market movement and macroeconomic variables. Although all estimation methods provide almost similar results, it can be considered that the proposed procedure is able to produce reliable results towards the presence of multicollinearity and outliers in the real data problem.

Acknowledgment

The authors are grateful for the financial support from (RAGS-FST-50214-55) from Universiti Sains Islam Malaysia.

References

[1] Hoerl A E and Kennard R W 1970 Ridge regression: applications to nonorthogonal problems *Technometrics* **12(1)** pp 69-82
[2] Hoerl A E and Kennard R W 1970 Ridge regression: biased estimation for nonorthogonal problems *Technometrics* **12(1)** pp 55-67
[3] Bagheri A and Midi H 2009 Robust estimations as a remedy for multicollinearity caused by multiple high leverage points *J. Math. Stat.* **5(4)** pp 311-21
[4] Lawless J F and Wang P 2010 A simulation study of ridge and other regression estimators *Comm. Statist. Theory Methods* **5(4)** pp 307-23
[5] Kibria B G 2003 Performance of some new ridge regression estimators *Comm. Statist. Simulation Comput.* **32(2)** pp 419-35
[6] Alkhamisi M, Khalaf G and Shukur G 2006 Some modifications for choosing ridge parameters *Comm. Statist. Theory Methods* **35(11)** pp 2005-20
[7] Alkhamisi M and Shukur G 2008 Developing ridge parameters for SUR model *Comm. Statist. Theory Methods* **37(4)** pp 544-64

[8] Muniz G and Kibria B G 2009 On some ridge regression estimators: an empirical comparison *Comm. Statist. Simulation Comput.* **38(3)** pp 621-30
[9] Mansson K, Shukur G and Kibria B G 2010 A simulation study of some ridge regression estimators under different distributional assumptions *Comm. Statist. Simulation Comput.* **39(8)** pp 1639-70
[10] Duzan H and Md. Shariff N S 2016 Solution to the multicollinearity problem by adding some constant to the diagonal *J. Mod. Appl. Stat. Methods* **15(1)** pp 752-73