Pak.j.stat.oper.res. Vol.19 No. 1 2023 pp 103-113 DOI: http://dx.doi.org/10.18187/pjsor.v19i1.4214

Pakistan Journal of Statistics and Operation Research

Prediction of KLCI Index Through Economic LASSO Regression Model and Model Averaging

Khuneswari Gopal Pillay[1*], Soh Pei Lin[2]

* Corresponding Author

1. Universiti Tun Hussein Onn, Malaysia, khuneswari@uthm.edu.my
2. Universiti Tun Hussein Onn, Malaysia, aw180252@siswa.uthm.edu.my

Abstract

The Financial Times Stock Exchange (FTSE) Bursa Malaysia KLCI Index is a key component in the development of Malaysia's economic growth and the complexity in terms of identifying the factors that have a substantial impact on the Malaysian stock market has always been a contentious issue. In this study, the macroeconomic factors of exchange rate, interest rate, gold price, consumer price index, money supply M1, M2, and M3, industrial production, and oil price were discussed by using economic LASSO regression and Bayesian Model Averaging (BMA) with monthly average and monthly end time-series data spanning from January 2015 to June 2021, with a total of 78 observations by using the R Studio. The findings demonstrate that month-end data is better suited for stock market prediction than month-average data and that the BMA model is more suitable than the LASSO model, as seen by lower Mean Square Error of Prediction, MSE(P) and Residual Mean Square Error of Prediction, RMSE(P) values. The exchange rate, gold price, and money supply have a negative association with the dependent variables, while the consumer price index has a positive relationship associated with the dependent variables. The consumer price index is the most significant contributing factor, whereas gold price is the least significant. The result depicted that the KLCI index has no significant relationship with the variables interest rate, money supply M2, M1, industrial production index, and oil price. In conclusion, investors could specifically focus on the positive contributor and put lesser attention on improving their portfolio return.

Key Words: KLCI, LASSO regression, Bayesian Model Averaging, Mean Square Error of Prediction, Residual Mean Square Error of Prediction.

1. Introduction

In Malaysia, Financial Times Stock Exchange (FTSE) Bursa Malaysia KLCI Index is known as a primitive stock market that appears local and includes the top 30 firms on Bursa Malaysia's Main Board in terms of full market capitalization (by Kenton. (2021)). Investors today tried to avoid their property from depreciating due to inflation, but they are still afraid of losing money by trading in the stock market. Finding a better forecast for the KLCI index by analysing its relationship with several variables will assist investors in making a more knowledgeable and secure decision when selecting a stock to invest in (by Murthy *et al*. (2017)).

Many researchers conclude that the movement of the KLCI Index is impacted by a variety of economic variables such as changes in macroeconomic determinants, changes in national policy, and even other illogical causes and they are attempting to determine whether economic conditions contribute to the volatility of KLCI Index (by Alzaid. (2016)). So, studying these markets and the factors that influence them can yield useful results for planning and achieving the specified goals for related industries such as investors, politicians, and the government of Malaysia (by Nasir *et al*. (2017)).

In this study, economic LASSO regression and Bayesian Model Averaging are used to find the macroeconomic factors, which are exchange rate, interest rate, gold price, consumer price index, money supply M1, M2, and M3, industrial production, and oil price was chosen for significant contribution to the prediction of KLCI index. The money supply is referred to the quantity of cash or currency circulating in an economy while M1 includes monies that are easily available for expenditure, M2 includes the money supply, M1 plus savings deposits and the M3 includes the money supply of M2 plus long-time deposits, institutional money-market funds, short-term repurchase agreements (by *The Fed.* (2015)).

A previous study proposed that to do research, should be done in a broader scope in terms of economic variables (by Zhao. (2010)). Hence, the study employs two types of data which are monthly-average (Case 1) and monthly-end (Case 2) time-series data spanning from January 2015 to June 2021, with a total of 78 observations. This research utilizes lasso regression and model averaging as a new method in stock market analysis to discover a broader body of knowledge. The main objective of this study is to identify the overall trend of the KLCI index, to compare the performance accuracy of prediction of the KLCI index through the LASSO Regression Model and Bayesian Model Averaging and finally to identify the significant macroeconomic variables that contribute to the prediction of KLCI index using the best-selected model.

2. Methodology

This section has a few sections that are covered, including the description and visualization of the data used, as well as the data pre-processing measures that are used to determine the existence of outliers and multicollinearity. Aside from that, the following sections would cover approaches such as LASSO regression, and model averaging.

2.1. Data Description

Table 1 shows the descriptions for each variable involved in the current study with the difference between Case 1 and Case 2. The KLCI index, which was derived from Yahoo Finance, was the primary dataset used in this analysis as the dependent variable. The dataset employed in this research is secondary data, with a total of 15 variables and 78 observations in total. There are two types of datasets divided within the study which are the monthly market average price dataset (Case 1) and the month-end closing price (Case 2) for the macroeconomic variables used from January 2015 to Jun 2021.

Table 1. Description of each variable in the study

Variables	Description	Case 1 Variables	Case 2 Variables
$\boldsymbol{y_1}$	KLCI Index Month Average	/	
$\boldsymbol{y_2}$	KLCI Index Month End		/
$\boldsymbol{x_1}$	Exchange Rate Month Average	/	
$\boldsymbol{x_2}$	Exchange Rate Month End		/
$\boldsymbol{x_3}$	Interest Rate Month Average	/	
$\boldsymbol{x_4}$	Interest Rate Month End		/
$\boldsymbol{x_5}$	Gold Price Month Average	/	
$\boldsymbol{x_6}$	Gold Price Month End		/
$\boldsymbol{x_7}$	Consumer Price Index	/	/
$\boldsymbol{x_8}$	M3	/	/
$\boldsymbol{x_9}$	M2	/	/
$\boldsymbol{x_{10}}$	M1	/	/
$\boldsymbol{x_{11}}$	Industrial Production Index	/	/
$\boldsymbol{x_{12}}$	Brent Oil Price Month Average	/	
$\boldsymbol{x_{13}}$	Brent Oil Month End		/

2.2. Outlier Detection

The presence of an outlier may affect the mean and variability, as well as the ultimate output of the model (by Walfish. (2006)). The goal is to remove any outliers from the data that might degrade the quality of the present study. The box-and-whisker plot, Cook's Distance, leverage statistics and DFFITS are utilized in the present investigation to find outliers in the dataset.

Box-and-Whiskers Plot is also known as a box plot. A box plot depicts the lower quartile (Q_1), upper quartile (Q_3), the median, lowest value, and maximum value in the box plot (by Laurikkala *et al.* (2000)). Meanwhile, Cook's distance is a measure generated for a certain regression model that is only impacted by the x variables within the dataset of the model (by Zhu, Ibrahim & Cho. (2012)). In general, outliers are defined as observations with a Cook's distance having more than four times the mean. The leverage point is characterised as an observation of extreme importance on predictor variables that did not affect the entire regression model but did influence the mean and variance of the dataset (by Chen. (2002)). The value of leverage varies between 0 and 1. The points are called influential points if the leverage values are greater than $3(\frac{p}{n})$ (by Kannan & Manoj. (2015)). Finally, DFFITS can be used to calculate the point where two variance estimators' equipped values differ. The conditions of DFFITS values of 1 or greater in limited samples of datasets, or DFFITS values greater than $2\sqrt{\frac{p}{n}}$ in broad samples of datasets are considered influential points (by Alkasadi *et al*. (2019)).

2.3. Multicollinearity

Multicollinearity occurs when the association between two or more explanatory variables is strongly linear. When there existed multicollinearity between two or more independent variables, the model interpretation would be difficult (by Assaf, Tsionas & Tasiopoulos. (2019)). This is because a researcher's study with high multicollinearity may result in an unstable and high error (by Senaviratna & Cooray. (2015)). As a result, the Pearson Correlation Matrix, and the Variance Inflation Factor (VIF) are utilized to discover multicollinearity in the current study's dataset and used in the explanation of the final best prediction model.

A correlation matrix is a table that displays the correlation coefficients for distinct variables in a symmetric matrix format. The table will show the correlations between all the possible ways in paring of the values. Pearson's correlation coefficient is the most often used correlation coefficient (by Stephanie. (2016)).

VIF is a handy method for assessing how much multicollinearity and frequency are in a series of variables. The higher the VIF, the greater the collinearity between the associated variables. The VIF is determined by the linear interaction of the predictors with the other independent variables (by Vu, Muttaqi & Agalgaonkar. (2015)). The value of VIF equals 1 demonstrates that the variables within the dataset do not have any multicollinearity. If multicollinearity is present, the VIF will be increased to a value greater than 1. VIFs of 1 to 5 suggest a weak relationship, while VIFs greater than 5 imply that the model's coefficients were incorrectly calculated (by Kim. (2019)).

2.4. K-fold Cross-Validation

In resampling techniques, K-fold cross-validation is characterized as the technique that most researchers prefer. The researcher defines it as "easy, accurate, and reliable." The model has a model section as well as classifier error estimation (by Anguita *et al.* (2009)). The first cross-validation of the macroeconomic dataset is performed by dividing the observations into an 80 % training set and a 20% testing set.

2.5. Least Absolute Shrinkage and Selection Operator (LASSO) Regression

The LASSO approaches are appealing and preferred by many researchers because they make it simple to define important independent variables and exclude unnecessary predictors (by Lee *et al.* (2021)). LASSO regression employs the L1 regularization method and the formula as in Equation (1).

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\beta_j \times x_{ij})^2 + \lambda \sum_{j=0}^{p}|\beta_j| \tag{1}$$

with the condition where some of $t > 0$, $\sum_{j=0}^{p}|\beta_j| < t$, where λ denotes the amount of shrinkage, $\sum_{i=1}^{n}(y_i - \sum_{j=0}^{p}\beta_j \times x_{ij})^2$ denotes the residual sum of squares and $\sum_{j=0}^{p}|\beta_j|$ denotes the sum of the absolute value of the magnitude of coefficients.

2.6. Bayesian Model Averaging

Model Averaging (MA) is the method of utilizing many models at once to make forecasts or infer parameters (by Khuneswari. (2015)). Bayesian Model Averaging (BMA) is going to be used in this research. BMA was first, a useful approach for overcoming the situation where model uncertainty is large and interferes with the prediction accuracy (by Hoeting *et al.* (1999)). BMA was suggested as a useful approach for overcoming the situation where model uncertainty is large and interferes with prediction accuracy. The model averaging would reduce the estimates of the weaker variables and choose the predictor models with the highest posterior probability (by Raftery, Madigan & Hoeting. (1997)). Assume there are M nominee models, $M = \{M_1, M_2, \ldots, M_k\}$ denoting the set of all models under consideration and Δ denoting the quantity of interest. The posterior distribution of the model Δ is given data D as in Equation (2).

$$\Pr(\Delta|\mathrm{D}) = \sum_{k=1}^{K} \Pr(\Delta|M_k, D)\Pr(M_k|\mathrm{D}) \tag{2}$$

where $Pr\ (M_k|D)$ represents the posterior probability of the model M_k and the value $k = 2^p - 1$.

2.7. Goodness-of-Fit Test

The goodness-of-fit test is a method for determining how well a sample of data fits into a population distribution using a normal distribution. In other words, the goodness-of-fit test is used to establish whether a variable is likely to originate from a previously selected distribution (by Jitkrittum *et al.* (2017)). The Kolmogorov-Smirnov hypothesis test is constructed as follows

H_0 : Sample data come from the same distribution

H_1 : Sample data come from different distribution

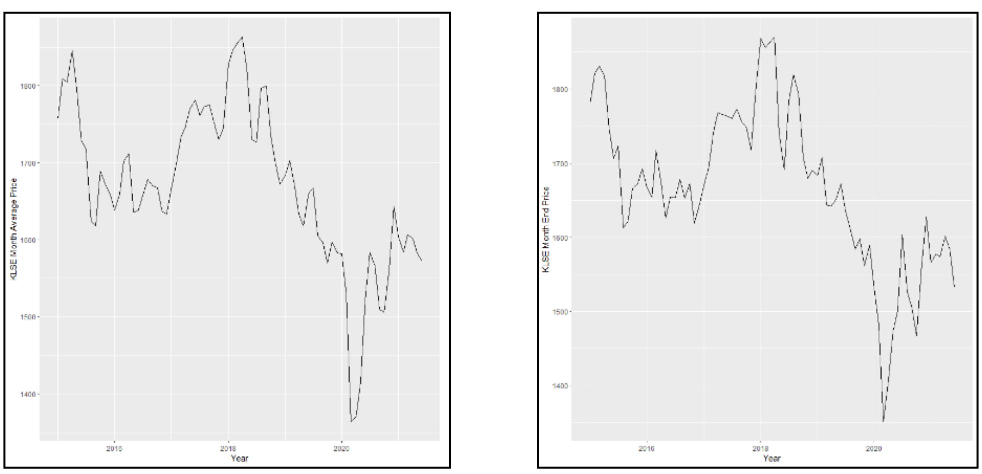
The Kolmogorov-Smirnov test is a simple calculation with no sample size or distribution restrictions (by Justel, Peña & Zamar. (1997)). The Kolmogorov-Smirnov test, QQ-Plot and the residual plot are utilized in the present goodness-of-fit test.

3. Results and Discussion

This section covers the analytical results that were discovered during the analysis with the beginning of the general trend distribution around the KLCI Index. This section also demonstrates the model-building method for the LASSO regression model and the Bayesian Model Averaging model, as well as how to overcome certain challenges such as influential points and non-normality concerns.

3.1. Data Visualization

The distributions of the KLCI Month Average Price (y_1) variable and KLCI Month End Price (y_2) variables are evaluated with a basic graphical analysis. Figure 1 depicts a time-series graph of both variables beginning on January 1, 2015 and ending on June 30, 2021. Both time series plots depict a similar trend and are on an overall downward trend. It began to manifest in 2015 as the trend began to decline until it became steady from 2016 to 2017 and it began to decline until it reached a minimal point in 2020 since the Covid-19 pandemic occurred in that year. Overall, the KLCI month average price time series plot is smoother than the KLCI month-end price plot in the visualisation.

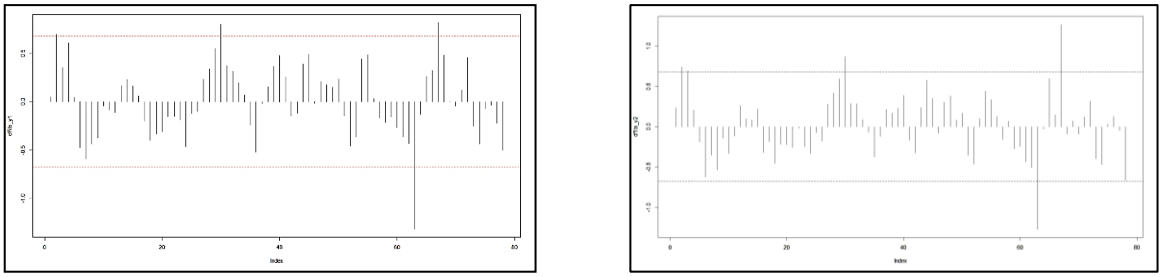


(a) (b)

Figure 1. Time series plot of KLCI (a) month average price and (b) month-end price from January 2015 until June 2021

3.2. Removal of Influential Observation

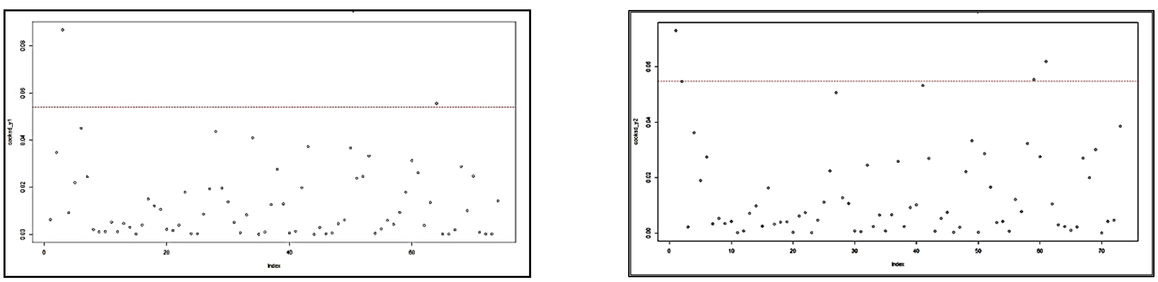
DFFITS, Cook's Distance, Leverage Statistics and Boxplot is applied to the Case 1 and Case 2 dataset. For DFFITS analysis, 2nd, 30th, 63rd, and 67th observations were identified as influential points for Case 1 while 2nd, 3rd, 30th, 63rd, and 67th observations were identified as influential points for Case 2 in Figure 2 and were eliminated before moving on to the next phase.



(a) (b)

Figure 2. DFFITS plot for (a) Case 1 and (b) Case 2

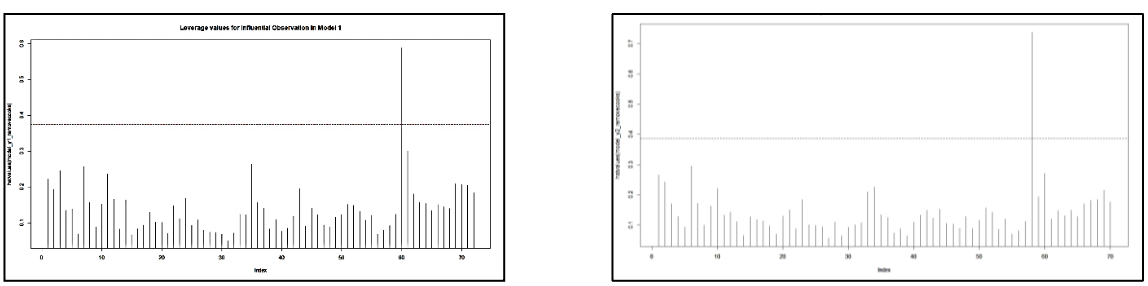
According to Figure 3, for the Cook's Distance method, 4th and 68th observations were identified as influential points in Case 1 while 1st, 62nd, and 65th observations were identified as influential points in Case 2 and were eliminated before moving on to the next phase.



(a) (b)

Figure 3. Cook's distance plot for (a) Case 1 and (b) Case 2

According to Figure 4, for the leverage statistics, the 64th observation has been identified as an influential point in Case 1 and Case 2 and has been deleted, proceeding into the next phase.



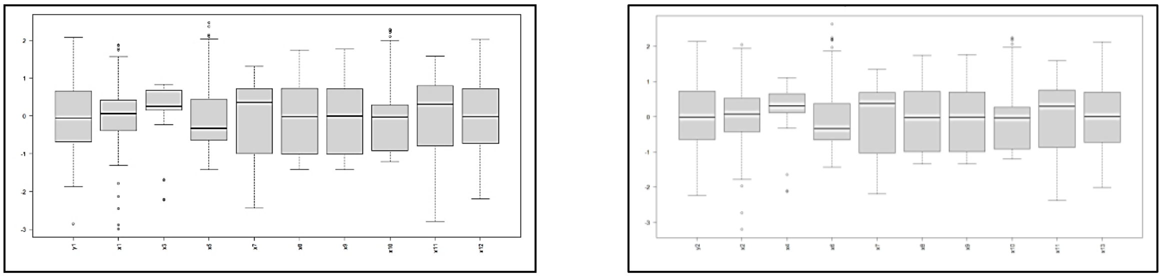
(a) (b)

Figure 4. Leverage statistic plot for (a) Case 1 and (b) Case 2

The boxplot for each standardized variable in Figure 5 was used to locate outliers in the remaining data. Several outliers were found in the KLCI Month Average Price (y_1), Exchange Rate Average (x_1), Interest Rate Average (x_3), Gold Price Average (x_5), and M1 (x_{10}) variables in Case 1, while Exchange Rate Average (x_2), Interest Rate Month-End (x_4), Gold Price Average (x_5), and M1(x_{10}) variables in Case 2 as seen in Figure 5. The boxplot's outliers are

 DOI: http://dx.doi.org/10.18187/pjsor.v19i1.4214

then detected and deleted. The total number of observations for the 10 variables has now been reduced from 78 to 47 for Case 1 and Case 2.

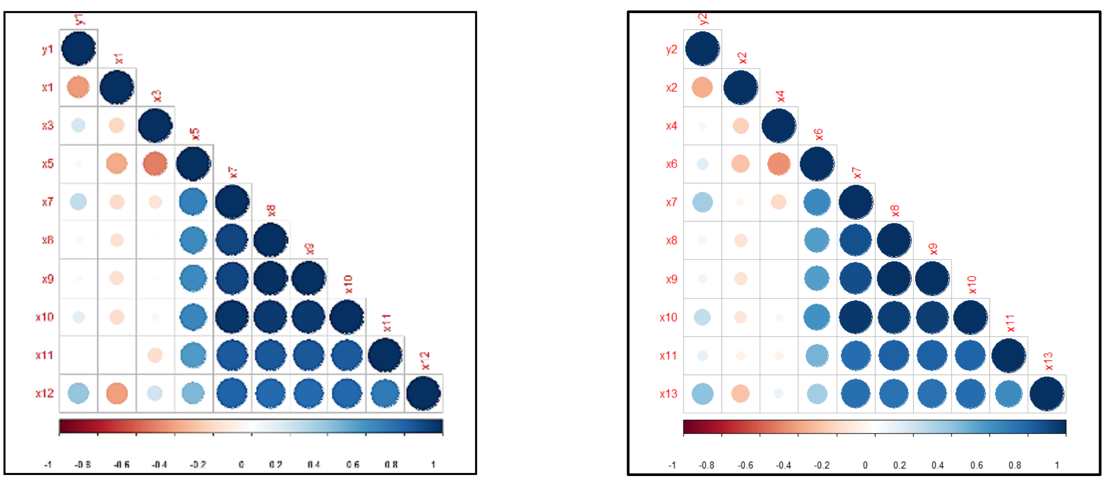


(a) (b)

Figure 5. Boxplot for (a) Case 1 and (b) Case 2

3.3. Multicollinearity

For the variables in the regression analysis for Case 1 and Case 2, a Pearson Correlation Matrix is shown to explain the relationship between them. According to Figure 6, the variable Exchange Rate Average (x_1) has a weak correlation with other variables while the variables Consumer Price Index (x_7), M3 (x_8), M2 (x_9) M1 (x_{10}) and Industrial Production (x_{11}) having a substantial correlation with most of the other variables in Case 1, while the variable Gold Price Month-End (x_4) having a weak correlation with other descriptive statistics while the variables Consumer Price Index (x_7), M3 (x_8), M2 (x_9) M1 (x_{10}) and Industrial Production (x_{11}) having a substantial positive correlation with most of the other variables.



(a) (b)

Figure 6. Pearson correlation matrix for (a) Case 1 and (b) Case 2

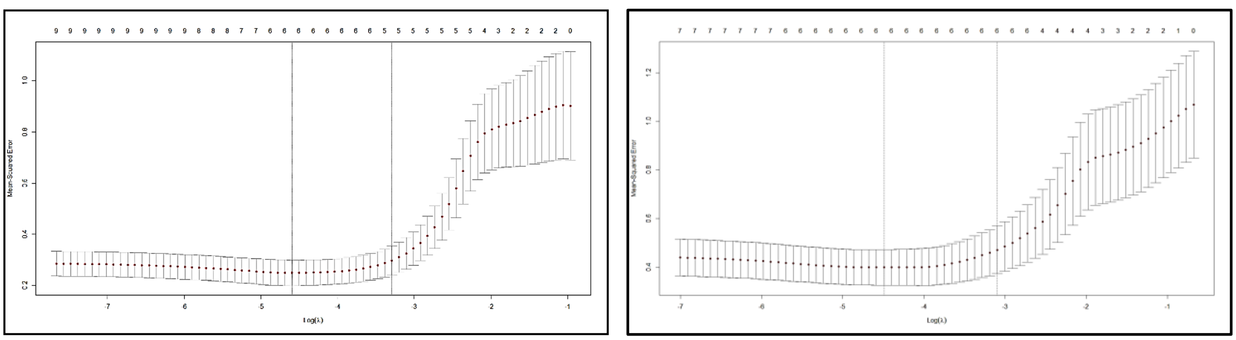
The VIF test is used to discover whether there is an issue with multicollinearity within the variables. Table 2 shows the results of the VIF test, were four variables, Consumer Price Index (x_7), M3 (x_8), M2 (x_9) and M1 (x_{10}), exist with substantial multicollinearity concerns, resulting in a VIF score of more than 10. However, the most significant benefit of the LASSO Regression and Bayesian Model Averaging Model is that both approaches can handle data with multicollinearity issues. Therefore, the variables with high multicollinearity are utilised in the stage of explanation of the results of both regression models.

Table 2. Description of each variable in the study for Case 1 and Case 2

Variables	Description	VIF value	
		Case 1	Case 2
x_1	Exchange Rate Month Average	3.3811	-
x_2	Exchange Rate Month End	-	3.3811
x_3	Interest Rate Month Average	3.8936	-
x_4	Interest Rate Month End	-	3.8936
x_5	Gold Price Month Average	7.0018	-
x_6	Gold Price Month End	-	7.0018
x_7	Consumer Price Index	44.1044	44.1044
x_8	M3	18598.7716	18598.7716
x_9	M2	19275.4479	19275.4479
x_{10}	M1	40.4936	40.4936
x_{11}	Industrial Production Index	5.2243	5.2243
x_{12}	Brent Oil Price Month Average	6.3558	-
x_{13}	Brent Oil Price Month End	-	6.3558

3.4. LASSO Regression

The LASSO model's cross-validated Mean-Squared Error (MSE) was displayed as shown in Figure 7. The minimal log lambda value is 0.01011, which is functioned as the best lambda value of the LASSO regression model. As indicated in Table 3, the optimal LASSO model with significant variables is obtained by employing the best lambda where the variables money supply M2 (x_9), M1 (x_{10}) and industrial production index (x_{11}) showed no significant relationship to the KLCI index for Case 1 and Case 2. According to Table 4, both Case 1 and Case 1 showed with MSE(P) value of 0.4725 and the RMSE(P) value of 0.6874 while Case 2 showed with MSE(P) value of 0.4182 and the RMSE(P) value of 0.6467.



(a) (b)

Figure 7. Cross-Validated MSE plot of LASSO (a) Case 1 and (b) Case 2

 DOI: http://dx.doi.org/10.18187/pjsor.v19i1.4214

Table 3. Significant coefficients variables for LASSO Case 1 and Case 2

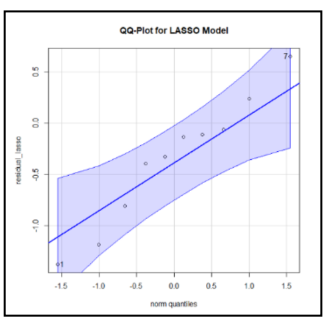
Variables	Variable Name	Coefficient	
		Case 1	Case 2
	Constant	0.0743	0.0523
x_1	Exchange Rate Month Average	-0.4798	-
x_2	Exchange Rate Month End	-	-0.4775
x_3	Interest Rate Month Average	0.0408	-
x_4	Interest Rate Month End	-	0.0851
x_5	Gold Price Month Average	-0.4669	-
x_6	Gold Price Month End	-	-0.2492
x_7	Consumer Price Index	1.6985	1.3653
x_8	M3	-1.5172	-1.3214
x_9	M2	-	-
x_{10}	M1	-	-
x_{11}	Industrial Production Index	-	-
x_{12}	Brent Oil Price Month Average	0.1671	-
x_{13}	Brent Oil Price Month End	-	0.3371

Table 4. MSE(P) and RMSE(P) for LASSO Case 1

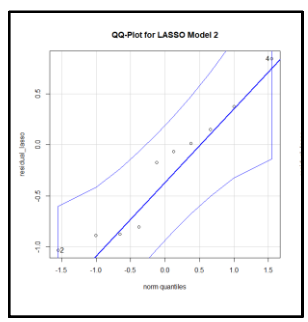
Evaluation Measure	Case 1	Case 2
MSE(P)	0.4725	0.4183
RMSE(P)	0.6874	0.6467

3.5. Goodness-of-Fit Test for LASSO Case 1

Figure 8 and 9 shows that all the data points in the QQ-plot are inside the blue reference point and that all the data points in the residual plot are randomly scattered about for Case 1 and Case 2. The p-value for the Kolmogorov-Smirnov test is more than 0.05 (refer to Table 5). As a result, the test data in Case 1 and Case 2 are well-fitting, and the residuals are normally distributed. Hence, the model is statistically accepted.

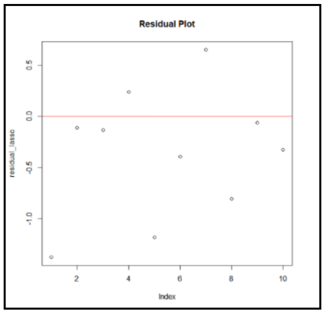


(a)

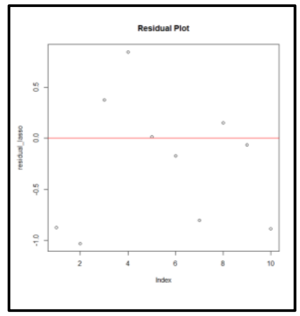


(b)

Figure 8. QQ-plot of LASSO (a) Case 1 and (b) Case 2



(a)



(b)

Figure 9. Residual plot of LASSO (a) Case 1 and (b) Case 2

Table 5. Significant coefficients variables for LASSO Case 1

Test	p-value	
	Case 1	Case 2
Kolmogorov-Smirnov	0.9500	0.7933

3.6. Bayesian Model Averaging

The model building procedure was conducted using the posterior probability as the weight for the models. The first model was chosen as the best projected BMA model with the highest posterior probability of 0.260 for Case 1 and 0.108 for Case 2. Table 6 shows the best projected BMA model with significant variables where the variables interest rate month average (x_3), money supply M3 (x_8), M1 (x_{10}), industrial production index (x_{11}), and oil price (x_{12}) show no significant relationship to the KLCI index for Case 1 and for Case 2, interest rate month end (x_4), M2 (x_9), M1 (x_{10}), industrial production index (x_{11}) and brent oil price month end (x_{13}) show no significant relationship to the KLCI index.

Table 6. Significant coefficients variables for BMA Case 1

Variables	Variable Name	Coefficient	
		Case 1	Case 2
	Constant	0.0773	0.0583
x_1	Exchange Rate Month Average	-0.5975	-
x_2	Exchange Rate Month End	-	-0.6422
x_3	Interest Rate Month Average	-	-
x_4	Interest Rate Month End	-	-
x_5	Gold Price Month Average	-0.6211	-
x_6	Gold Price Month End	-	-0.4716
x_7	Consumer Price Index	2.0128	1.7352
x_8	M3	-	-1.3145
x_9	M2	-1.6079	-
x_{10}	M1	-	-
x_{11}	Industrial Production Index	-	-
x_{12}	Brent Oil Price Month Average	-	-
x_{13}	Brent Oil Price Month End	-	-

The MSE(P) and RMSE(P) are obtained by comparing the remaining 20% of the testing test using the best prediction BMA model. Therefore, referring to Table 7, the MSE(P) value is 0.5495, while the RMSE(P) value is 0.7413 for Case 1 and the MSE(P) value is 0.3731, while the RMSE(P) value is 0.6108 for Case 2.

Table 7. MSE(P) and RMSE(P) for BMA Case 1 and Case 2

Evaluation Measure	Case 1	Case 2
MSE(P)	0.5495	0.3731

RMSE(P)	0.7413	0.6108

3.7. Comparison Between Model-Building Approaches

Model-building on the month-average dataset (Case 1) and month-end dataset (Case 2) were conducted. Table 8 shows the two types of datasets (Case 1 and Case 2) approaches with a total of 4 models. Only one best model with the smallest MSE(P) is chosen.

Table 8. Comparison between LASSO model and BMA model in Case 1 and Case 2

Model	Method	Number of Significant Variables	MSE(P)	RMSE(P)
Case 1	LASSO	6 out of 9 variables	0.4725	0.6874
	BMA	4 out of 9 variables	0.5495	0.7413
Case 2	LASSO	6 out of 9 variables	0.4182	0.6467
	BMA	4 out of 9 variables	0.3731	0.6108

The number of significant variables for each case and method is shown in Table 3 and Table 8. According to Table 8, the best model is Case 2 which uses the BMA regression method. The best model could be written as in Equation (3), where the Exchange Rate Month-End (x_2), Gold Price Month-End (x_6), Consumer Price Index (x_7), and M3(x_8) are the factors that have a substantial impact on the KLCI index.

$$\hat{Y} = 0.0583 - 0.6422\, exchange\ rate - 0.4716\, gold\ price + 1.7352\, CPI - 1.3145\, M3 \tag{3}$$

The best BMA regression model was then used to forecast the KLCI month-end values for the following three months, July, August, and September in 2020. As can be seen in Table 9, which compares the real value and the predicted values using the best regression model equation, the value for July 2021 is remarkably close when the time is nearer, indicating that the regression model is suitable for short-term prediction.

Table 9. Prediction by using the best model

Time	Original value	Prediction value
July 2021	1494.60	1501.937
August 2021	1601.38	1550.260
September 2021	1537.80	1589.950

4. Conclusion

For the first objective, a conclusion of the overall KLCI month average and month-end index have both shown a progressively dropping trend in Malaysia between the year 2015 and the year 2021. This revealed that Malaysia's economic situation was deteriorating which should be considered by the nation and policymakers.

For the second objective, the current study has provided that the month-end variables are more suitable than using month-average variables because the overall MSE(P) and RMSE(P) of both LASSO and BMA models showed the lowest value in using the month-end index dataset. Also, overall, the current study has proved that BMA regression is the best technique used in the prediction of the KLCI index compared to the LASSO regression.

In the third objective, by using the best model selected, the significant macroeconomic variables are exchange rate, gold price, consumer price index and M3. The consumer price index is the most significant contributing factor, whereas gold price is the least significant. The study found that the KLCI index has no significant relationship with the variables interest rate, money supply M2, M1, industrial production index, and oil price.

Overall, all the objectives for the study were met, and significant contributions could be defined in depth. It was recommended that future studies strive to investigate a broader variety of variables. Also, since there are influential points and outliers left after employing the DFFITS, Cook's Distance, and leverage values, additional methods, such as DFBETAS.

References

1. Kenton, W. (2021, December 1). *Kuala Lumpur Stock Exchange (KLCI) Definition.* Investopedia. Retrieved from https://www.investopedia.com/terms/k/KLCI.asp

2. Murthy, U., Anthony, P. & Vighnesvaran, R. (2017). Factors Affecting Kuala Lumpur Composite Index (KLCI) Stock Market Return in Malaysia. *International Journal of Business and Management, 12*(1), 122-132.
3. Alzaid, S. (2016). The Kuala Lumpur stock exchange composite index (KLCI CI) and economic forces. *South-East Asia Journal of Contemporary Business, Economics and Law, 10*(3), 53-64.
4. Nasir, N. M., Hassan, N. M., Nasir, Z. A. & Harun, M. F. M. (2017). Macroeconomic Factors as the Determinants of Stock Market Return in Malaysia: Multivariate Cointegration and Causality Analysis. *Terengganu International Finance and Economics Journal (TIFEJ), 3*(1), 38-49.
5. *The Fed - What is the money supply? Is it important?* (2015, December 16). Federal Reserve System. Retrieved from https://www.federalreserve.gov/faqs/money_12845.html
6. Zhao, Y. & Bondell, H. (2020). Solution paths for the generalized lasso with applications to spatially varying coefficients regression. *Computational Statistics & Data Analysis, 142,* 106821.
7. Walfish, S. (2006). A review of statistical outlier methods. *Pharmaceutical Technology, 30*(11), 82–86.
8. Laurikkala, J., Juhola, M., Kentala, E., Lavrac, N., Miksch, S. & Kavsek, B. (2000, August). Informal identification of outliers in medical data. In *Fifth international workshop on intelligent data analysis in medicine and pharmacology* (Vol. 1, pp. 20-24).
9. Zhu, H., Ibrahim, J. G. & Cho, H. (2012). Perturbation and scaled Cook's distance. *Annals of statistics, 40*(2), 785.
10. Chen, C. (2002). Robust Regression and Outlier Detection with the ROBUSTREG procedure. *In Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*.
11. Kannan, K. S. & Manoj, K. (2015). Outlier detection in multivariate data. *Applied Mathematical Sciences, 47*(9), 2317-2324.
12. Alkasadi, N. A., Ibrahim, S. A. F. W. A. T. I., Abuzaid, A. H., Yusoff, M. I., Hamid, H., Zhe, L. W. & Abd Razak, A. (2019). Outlier Detection in Multiple Circular Regression Model using DFFITC Statistic. *Sains Malaysiana*, *48*(7), 1557-1563.
13. Assaf, A. G., Tsionas, M. & Tasiopoulos, A. (2019). Diagnosing and correcting the effects of multicollinearity: Bayesian implications of ridge regression. *Tourism Management*, 71, 1-8.
14. Senaviratna, N. A. M. R. & Cooray, T. M. J. A. (2019). Diagnosing multicollinearity of logistic regression model. *Asian Journal of Probability and Statistics, 5*(2), 1-9.
15. Stephanie, G. (2016, May 11). *Correlation Matrix: Definition.* Statistics How To. Retrieved from https://www.statisticshowto.com/correlation-matrix/
16. Vu, D. H., Muttaqi, K. M. & Agalgaonkar, A. P. (2015). A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. *Applied Energy, 140,* 385-394.
17. Kim, J. H. (2019). Multicollinearity and misleading statistical results. *Korean journal of anesthesiology, 72*(6), 558.
18. Anguita, D., Ghio, A., Ridella, S. & Sterpi, D. (2009). K-Fold Cross Validation for Error Rate Estimate in Support Vector Machines. *International Conference on Data Mining, June 2014,* 291–297.
19. Lee, J. H., Shi, Z. & Gao, Z. (2021). On LASSO for predictive regression. *Journal of Econometrics.*
20. Khuneswari, G. (2015). *Model selection and model averaging in the presence of missing values* (Doctoral dissertation, University of Glasgow).
21. Hoeting, J. A., Madigan, D., Raftery, A. E. & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical science,* 382-401.
22. Raftery, A. E., Madigan, D. & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*(437), 179-191.
23. Jitkrittum, W., Xu, W., Szabó, Z., Fukumizu, K. & Gretton, A. (2017). A linear-time kernel goodness-of-fit test. *Advances in Neural Information Processing Systems, 2017-December* (Nips), 262–271.
24. Justel, A., Peña, D. & Zamar, R. (1997). A multivariate Kolmogorov-Smirnov test of goodness of fit. *Statistics & Probability Letters, 35*(3), 251-259.