Enhanced Knowledge in Sciences and Technology Vol. 1 No. 2 (2021) 30-39

EKST

Homepage: http://publisher.uthm.edu.my/periodicals/index.php/ekst
e-ISSN : 2773-6385

Prediction of Unemployment Rate in Malaysia Based on Macroeconomic Factors

Khuneswari Gopal Pillay[1], **Teo Bi En**[2]

[1]Khuneswari Gopal Pillay,
Department Mathematics and Statistics,
Universiti Tun Hussein Onn Malaysia, Pagoh, 84600, MALAYSIA

[2]Teo Bi En,
Department Mathematics and Statistics,
Universiti Tun Hussein Onn Malaysia, Pagoh, 84600, MALAYSIA

*Corresponding Author Designation

DOI: https://doi.org/10.30880/ekst.2021.01.02.004
Received 11 May 2021; Accepted 16 June 2021; Available online 29 July 2021

Abstract: The unemployment rate has become a critical issue, not only in Malaysia but a worldwide phenomenon. Hence, the macroeconomic factors that significantly affect the unemployment rate in Malaysia were investigated in this research. The data used was obtained from Trading Economics and the Central Bank of Malaysia. At first, the influential points were detected and removed using Cook's Distance. The correlation and multicollinearity were then tested to investigate the relationships among the variables such as unemployment rate, gross domestic product (GDP) growth rate, inflation rate, foreign direct investment (FDI), population growth rate and exchange rate. The LASSO regression method was applied to identify the significant macroeconomic factors that affect the unemployment rate in Malaysia. Three different LASSO models were formed under different conditions, which included the model without data transformation (Model A), the model with data transformation (Model B) and the model with data transformation except for the population growth rate (Model C). In conclusion, Model A was chosen as the best LASSO model as it has the smallest value of MSE(P) compared to Model B and Model C. The inflation rate, FDI, population growth rate and exchange rate were the significant macroeconomic factors that causing an increment or decrement of the unemployment rate in our country. Therefore, fiscal, and monetary policy should be enforced by the government and policymakers to improve the issue of unemployment thus stabilising the economy of Malaysia.

Keywords: Unemployment Rate, Macroeconomic Factors, LASSO Regression, Data Transformation, Mean Square Error of Prediction

Corresponding author: khuneswari@uthm.edu.my.com

1. Introduction

Unemployment has become a critical issue worldwide including in Malaysia. According to the Department of Statistics Malaysia (DOSM), the unemployment rate in Malaysia was found that had decreased from 3.1% in 2013 to 2.9% in 2014 and increased to 3.4% in 2016. National Economic Action Council with the collaboration of the Department of Human Manpower also found that there were 59000 graduates and diploma holders who were unemployed, and 30000 graduates worked in the mismatched field with their higher educational qualifications in Malaysia [1].

In this research, several macroeconomic factors were adopted from previous research to extend the investigation, such as Gross Domestic Product (GDP), inflation rate, population, exchange rate and Foreign Direct Investment (FDI). The significant macroeconomic factors were then investigated and used in the prediction of the unemployment rate in Malaysia to provide insights and guidance for relevant parties to act accordingly.

The main objectives of this research are to investigate the relationship between the macroeconomic factors and the unemployment rate in Malaysia, to identify the significant macroeconomic factors that affect the unemployment rate in Malaysia, and to predict the unemployment rate in Malaysia based on macroeconomic factors. All the objectives were achieved by applying the LASSO regression method.

2. Methodology

2.1 Data Description

The data set was obtained and downloaded from several websites such as Trading Economics (https://tradingeconomics.com/) and Central Bank of Malaysia (https://www.bnm.gov.my/). The data set used in this research is quarterly time-series data from the year 2005 to 2019, and in a total of 60 samples of quarterly data for each variable. The unemployment rate in Malaysia is chosen to be the dependent variable. There is a total of 5 macroeconomic factors in this research. The factors selected as independent variables are including GDP growth rate, inflation rate, FDI, population growth rate and exchange rate.

2.2 Cook's Distance

An influential point can negatively affect the regression model, it must be examined and removed to improve data accuracy. Cook's Distance provides a way to identify the influential outliers in each set of variables [2]. The formula of Cook's Distance is

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{(p+1)\hat{\sigma}^2} \qquad Eq.1$$

where $\hat{Y}_j$is the j^{th} fitted dependent value, $\hat{Y}_{j(i)}$ is the j^{th} fitted dependent value where the fit does not include i^{th} observation, p is the number of coefficients in the regression model and $\hat{\sigma}^2$ is the estimated variance from the fitted value based on all observations. The value of any data point that exceeds the cut-off rate of 4/*n* can be considered as an influential point, where *n* is the sample size [3].

2.3 Correlation

Correlation between the data can be identified by using correlation analysis which the correlation coefficient between sets of variables are being calculated [4]. It is usually referred to as Pearson's product-moment correlation and the formula [5] is

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 - \sum(Y_i - \bar{Y})^2}} \qquad Eq.2$$

where r is the correlation coefficient, X_i is the value of the i^{th} variable X in a sample, $\bar{X}$ is the mean of the values of the variable X, Y_i is the value of the i^{th} variable Y in a sample and $\bar{Y}$ is the mean of the values of the variable Y. The r is resulted in the range from -1 to 1, where 1 indicates a strong positive relationship, -1 indicates a strong negative relationship while zero indicates no relationship at all. These values are then fed into the correlation matrix to visualise each of the variables' relationship between one another.

2.4 Multicollinearity

Multicollinearity is a case of multiple regression in which the independent variables are highly correlated with each other [6]. When performing any multiple regression, examination for the existence of multicollinearity within the data set should be carried out [7]. It can be detected and identified by using variance-inflation factors (VIF) [8]. The VIF for each variable can be calculated by

$$VIF_i = \frac{1}{1 - R_i^2} \qquad Eq.3$$

where R_i^2 is the coefficient of multiple determination obtained from regressing X_i on the other regressor variables. Variables with a VIF value that exceeds 10 should be removed [8].

2.5 Data Transformation

Data transformation is commonly used in regression analysis to improve the skewness of the data and obtain better prediction results [9]. There are several types of data transformation method and different types of transformation methods have different criteria. For example, reciprocal transformation can be only applied for non-zero values and the values are strict to be positive for Box-Cox transformation. In this research, the Johnson transformation is chosen to be applied for the non-normal variables, with the p-value < 0.05 that determined using the Anderson-Darling test.

2.6 Least Absolute Shrinkage and Selection Operator (LASSO) Regression

LASSO is a regression method that always used to enhance the ability and accuracy of prediction of the model it forms by performing variable selection and regularisation. The regularisation technique called L1-norm, which is the sum of the absolute coefficients to penalise and shrink the regression coefficients towards zero [10]. Cross-validation (CV) is used to split the collected data randomly into training data and testing data [11]. The optimal tuning parameter (λ) was chosen to obtain the most accurate prediction by building the best LASSO model. After that, the prediction is carried out using the test data with the LASSO model. Mean squared error of prediction (MSE(P)) is then used to measure how well the model predicts the dependent variable [12] by the formula

$$MSE(P) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{Y}_i - Y_i)^2 \qquad Eq.4$$

where i = 1,2, …,n. $\widehat{Y}_i$ is estimated Y of test values and Y_i is the actual test values used for prediction. The value of MSE(P) closer to zero, means that the prediction is closer to the true value.

2.7 Goodness-of-Fit Test

The goodness-of-fit test is used to test if the sample data fits a distribution from a population with a normal distribution. The Kolmogorov-Smirnov test, Anderson-Darling test and Shapiro-Wilk test are usually used as the goodness-of-fit test for a continuous distribution. These tests are typically run using software such as R software. The hypothesis test for the goodness-of-fit test is

Null hypothesis, $H_0 =$ The residuals, e_i are normally distributed.

Alternative hypothesis, $H_1 =$ The residuals, e_i are not normally distributed

Normally, the null hypothesis is rejected if the p-value < 0.05. Thus, it will be concluded as the residuals are not normally distributed. Besides the calculations, the data distribution can also be investigated using some diagnostic plots such as Q-Q (quantile-quantile) plot and residual plot [13].

2.8 Two-Sample T-Test

Two-sample t-test is usually applied in a comparison of means from two populations or groups, to investigate whether there is a significant difference between them. The hypothesis test for comparing the difference between two means is

$$\text{Null hypothesis, } H_0 = \mu_1 - \mu_2 = 0$$

$$\text{Alternative hypothesis, } H_1 = \mu_1 - \mu_2 \neq 0$$

where μ represents the mean value. The null hypothesis is concluded as there is no significant difference between the two means while the alternative hypothesis is concluded as there is a significant difference between the two means. With assuming the equal variances, pooled variance (s_p^2) is calculated to obtain a better estimate. The formula of the test statistic, pooled variance and degree of freedom (df) are

$$Test\ statistic, t = \frac{\mu_1 - \mu_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \qquad Eq.5$$

$$Pooled\ variance, s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \qquad Eq.6$$

$$Degree\ of\ freedom, df = n_1 + n_2 - 2 \qquad Eq.7$$

where s^2 is the variance of sample, n is the sample size and μ is the mean value. The null hypothesis is rejected if the test statistic, $|t| >$ critical value. The null hypothesis is also rejected if the p-value $\leq \alpha$, where the significance level, $\alpha = 0.05$, is the most used in hypothesis tests.

3. Results and Discussion

3.1 Variable Description

3.1.1 Cook's Distance

Figure 1 shows the bar plot of Cook's Distance that used to identify the influential points in this research. The influential point is identified by using the cut-off rate of $4/n$, which equal to 0.067. Based on Figure 1, there are 4 influential points clearly labelled with the numbers of data. As the influential points can impact the result and accuracy of a regression model, thus all the 4 influential points are removed. The remaining data are used for all the regression analysis in this research.

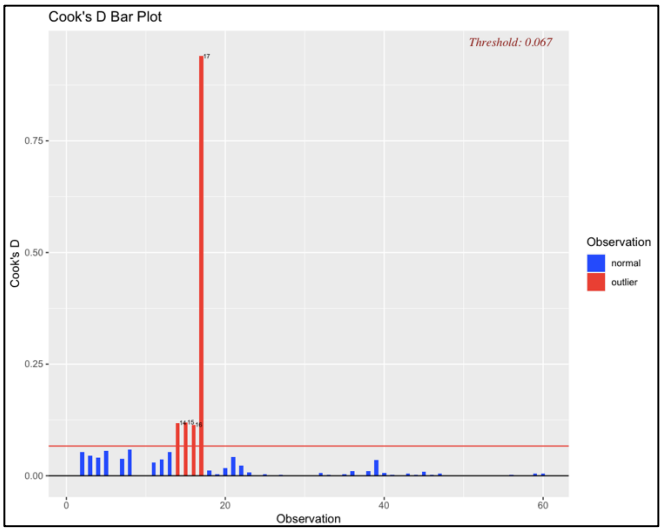

Figure 1: Cook's Distance bar plot

3.1.2 Correlation Matrix

Figure 2 shows the correlation matrix of the variables used in this research. The positive correlations are displayed in blue colour whereas the negative correlations in red colour. The colour intensity and the size of the circle are proportional to the correlation coefficients. Overall, the population growth rate and exchange rate are having positive correlations with the unemployment rate. The exchange rate has the strongest positive correlation with the unemployment rate while the FDI has the strongest negative correlation with the unemployment rate compared to other variables. The inflation rate has the weakest negative correlation due to the smallest size of the circle and the lightest colour intensity. The correlation between the inflation rate and the exchange rate has resulted in no correlation as there is no colour or circle exists.

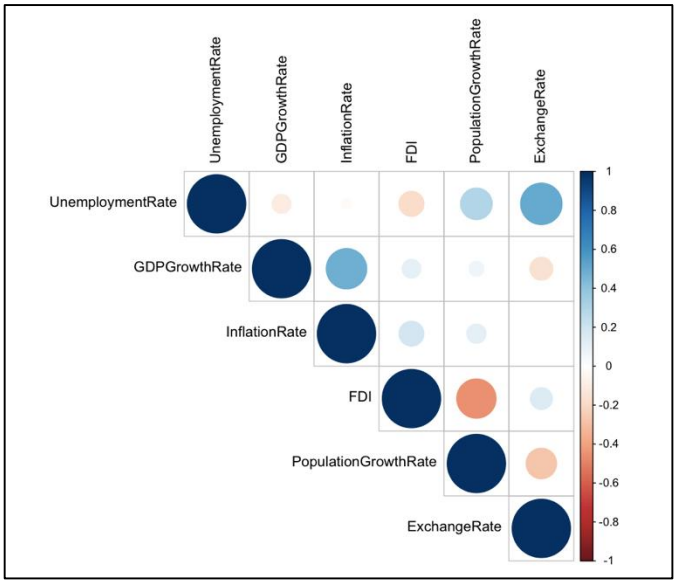

Figure 2: Correlation matrix of variables

3.1.3 Multicollinearity Test

Table 1 shows the values of VIF for all independent variables used in this research. Since all the independent variables with the values of VIF around 1, this means that there are no multicollinearity effects between the independent variables. Thus, there is no elimination occurs for the variables.

Table 1: Variance inflation factor (VIF)

Variable	Description	Value of VIF
X_1	GDP Growth Rate	1.3467
X_2	Inflation Rate	1.3934
X_3	FDI	1.3599
X_4	Population Growth Rate	1.4077
X_5	Exchange Rate	1.1187

3.1.4 Normality Test

Normality is a basic assumption for regression. Hence, the normality tests such as the Q-Q plot and Anderson-Darling test are applied in this research. Figure 3 shows the Q-Q plot for the model without influential points. The plot shows that the normality might not be met by this model. The data points should be scattered approximately along the reference line in the Q-Q plot.

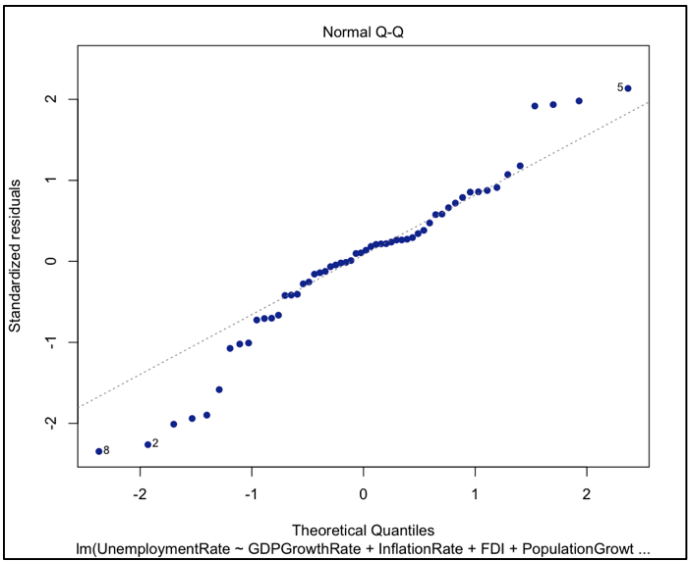

Figure 3: Q-Q plot for the regression model without influential points

Table 2 shows all the p-values of the Anderson-Darling test for the independent variables. Based on the result shown in Table 2, only two variables can be considered as normally distributed as their p-values > 0.05. The other variables with the p-value < 0.05 might need to undergo the process of data transformation before proceeding with the model-building process.

Table 2: *P*-value of the Anderson-Darling test for independent variables

Variable	Description	P-value
X_1	GDP Growth Rate	0.0003
X_2	Inflation Rate	0.1257
X_3	FDI	0.5039
X_4	Population Growth Rate	0.0037
X_5	Exchange Rate	0.0011

3.2 LASSO Model-Building without Data Transformation (Model A)

Model A is the LASSO regression model, which is built without any data transformation, to investigate whether there is a difference between the model with and without data transformation.

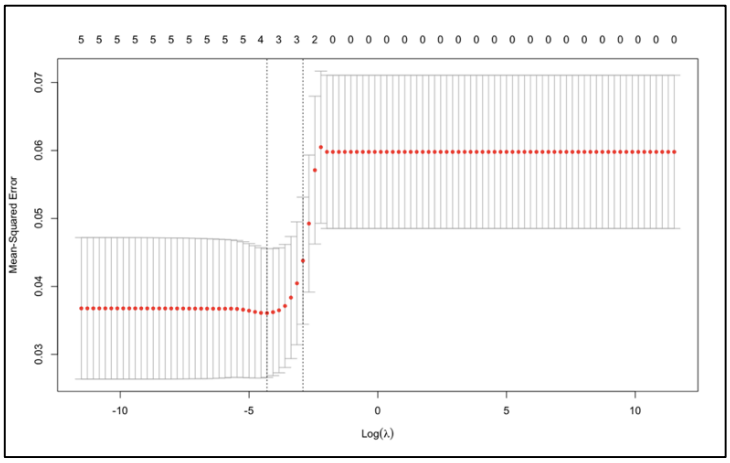

Figure 4: Cross-validated MSE plot of LASSO Model A

Figure 4 shows the minimum log(λ) value of the cross validated MSE plot. The optimal lambda value obtained from the plot is 0.0135. With the optimal lambda value, the best predicted LASSO Model A is formed with significant variables. The best-predicted Model A is,

Unemployment rate = 1.6167 – 0.0012*Inflation rate – 0.0044*FDI + 0.3988*Population growth rate + 0.2952*Exchange rate

3.3 LASSO Model-Building with Data Transformation (Model B)

Model B is the LASSO model, which is built with data transformation, with the purpose to determine whether there is a difference between the model with and without data transformation. Therefore, the Johnson transformation was carried out for the non-normal variables. After the data transformation, only the population growth rate remains not normally distributed.

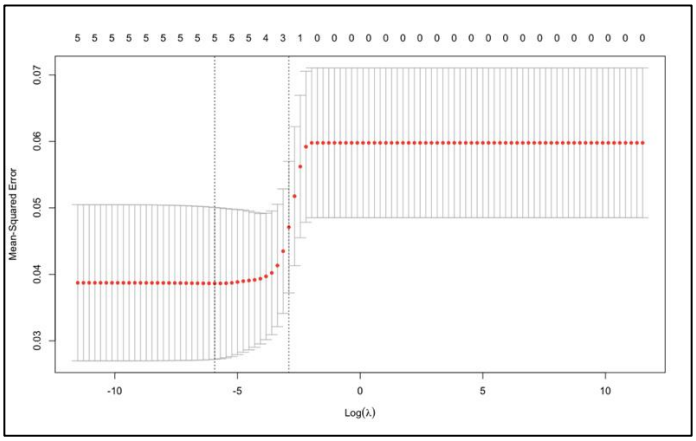

Figure 5: Cross-validated MSE plot of LASSO Model B

The optimal lambda value for Model B is 0.0027, which was obtained from the minimum log(λ) value of the cross validated MSE plot in Figure 5. With the optimal lambda value, the best predicted LASSO Model B was formed with significant variables as below.

Unemployment rate = 3.42 + 0.0257*GDP growth rate – 0.0348*Inflation rate – 0.0086*FDI + 0.0881*Population growth rate + 0.1682*Exchange rate

3.4 LASSO Model-Building with Data Transformation Except For Population Growth Rate (Model C)

Model C is the LASSO model, which is built with data transformation, except for the transformation of population growth rate, as it remains non-normal after data transformation. Therefore, the Johnson transformation was carried out without transforming the population growth rate. After the transformation, all the variables were normally distributed except the population growth rate.

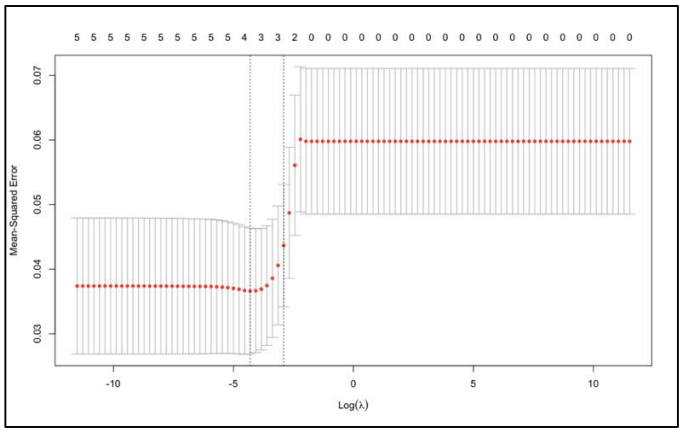

Figure 6: Cross-validated MSE plot of LASSO Model C

The optimal lambda value for Model C is 0.0135, which was obtained from the plot in Figure 6. With the optimal lambda value, the best predicted LASSO Model C was formed with significant variables. The best predicted LASSO Model C is

Unemployment rate = 2.7262 – 0.0007*Inflation rate – 0.0051*FDI + 0.3724*Population growth rate + 0.142*Exchange rate

3.5 Goodness-of-Fit Test of the Three Best Predicted LASSO Models

Table 3 shows the results of the normality tests for the three best LASSO models (A, B and C). The null hypothesis will not be rejected if the p-values > 0.05, thus the residuals will be concluded as normally distributed. Since the p-values for all three best models are greater than 0.05, the residuals can be concluded as normally distributed.

Table 3: Normality test for the three best models on testing data

Normality Test	*P*-value		
	Model A	Model B	Model C
Shapiro-Wilk	0.1734	0.5918	0.6136
Kolmogorov-Smirnov	0.7106	0.7682	0.9902
Anderson-Darling	0.1711	0.5058	0.7905

3.6 Comparison Between Three Best Predicted LASSO Models

3.6.1 Mean Squared Error of Prediction (MSE(P))

Based on the values of MSE(P) that resulted in Table 4, the best predicted LASSO Model A has the smallest value of error which is 0.0308 among all the three models. Therefore, Model A is chosen as the best LASSO model for the prediction. Model A is built without any data transformation and influential point, using the optimal lambda value of 0.0135. Hence, the significant macroeconomic factors that affect the unemployment rate are including the inflation rate, FDI, population growth rate and exchange rate.

Table 4: Significant variables, optimal lambda (λ) and MSE(P) of the three LASSO models (A, B and C)

Model	Significant Variables					Optimal λ	MSE(P)
	GDP growth rate	Inflation rate	FDI	Population growth rate	Exchange rate		
A		✓	✓	✓	✓	0.0135	0.0308
B	✓	✓	✓	✓	✓	0.0027	0.0429
C		✓	✓	✓	✓	0.0135	0.0336

3.6.2 Two-Sample *T*-Test

A hypothesis testing is carried out using the method of two-sample t-test to identify whether there is a significant difference among all the three LASSO models (A, B and C).

Table 5: Hypothesis testing for three sets of comparison

Set	Model	Hypothesis Testing
1	A and B	Null hypothesis, $H_0 = MSE(P)_A - MSE(P)_B = 0$ Alternative hypothesis,$H_1 = MSE(P)_A - MSE(P)_B \neq 0$
2	A and C	Null hypothesis, $H_0 = MSE(P)_A - MSE(P)_C = 0$ Alternative hypothesis,$H_1 = MSE(P)_A - MSE(P)_C \neq 0$
3	B and C	Null hypothesis,$H_0 = MSE(P)_B - MSE(P)_C = 0$ Alternative hypothesis,$H_1 = MSE(P)_B - MSE(P)_C \neq 0$

A total of three sets of comparison are carried out which included Model A and B (Set 1), Model A and C (Set 2) and for Model B and C (Set 3). The hypothesis testing and the results of the two-sample *t*-test for the three sets comparison are shown in Table 5 and Table 6, respectively.

Table 6: Results of the two-sample *t*-test for three sets of comparison

Set	Case	*t*-statistic	Critical *t*-value	*P*-value	Decision
1	A and B	-0.5415	2.0739	0.5936	Do not reject H_0
2	A and C	-0.1357	2.0739	0.8933	Do not reject H_0
3	B and C	0.4112	2.0739	0.6849	Do not reject H_0

A two-sample *t*-test with equal variance was carried out for the comparison since the sample size for all the three LASSO models were equal. Since all the test statistics, $|t|$ for all three sets, the comparison is smaller than the critical *t*-value, which means that test statistics were fell in the non-rejection region, thus the null hypothesis cannot be rejected. Therefore, the results can be concluded as there are no significant differences among the three best LASSO models.

4. Conclusion

In a nutshell, all the objectives were achieved by using LASSO regression, with building and choosing the best LASSO model for the prediction. Since the best predicted LASSO Model A formed with the smallest value of MSE(P) compared to the others, Model A is chosen as the best LASSO model for the prediction of the unemployment rate in Malaysia. Hence, the significant macroeconomic factors that affect the unemployment rate were identified in this research by using LASSO regression, which included the inflation rate, FDI, population growth rate and exchange rate.

From the equation of the best predicted LASSO Model A, the inflation rate and FDI were found as having a negative relationship with the unemployment rate whereas the population growth rate and exchange rate were concluded as having a positive relationship with the unemployment rate. Since the best LASSO model, which is Model A, was built without the data transformation, it means that that there might be not necessary for the data transformation in LASSO regression. There might be problems even when using some popular data transformations such as log transformation [14]. The result of this finding might give a direction for the government and policymakers to implement the fiscal and monetary policy for improving the unemployment issue in Malaysia.

Future researchers are recommended to use a larger sample size in future research for obtaining a better result. Besides, future researchers are also recommended to determine the minimum required sample size before implementing multiple linear regression for prediction purposes. The number of independent variables can be an important factor to determine the minimum required sample size [15]. Therefore, future researchers may study the other factors that might affect the unemployment rate in Malaysia.

Acknowledgment

The authors thank the Indonesian Ministry of Riset dan Technology for providing funding for this research. We were also thankful to the Indonesian Institute of Sciences for providing access and technical support (ELSA-LIPI).

References

[1] Z. Hanapi and M.S. Nordin, "Unemployment among Malaysia graduates: Graduates' attributes, lecturers' competency and quality of education," Procedia-Social and Behavioral Sciences, vol. 112, no. 2014, pp. 1056–1063, 2014.

[2] R. D. Cook, "Detection of Influential Observation in Linear Regression," Technometrics, vol. 19, no. 1, pp. 15–18, 1977.

[3] N. Altman and M. Krzywinski, "Analyzing outliers: influential or nuisance?" Nature Methods, vol. 13, no. 6, pp. 535, 2016.

[4] R. Taylor, "Interpretation of the correlation coefficient: a basic review," Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–39, 1990.

[5] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," The American Statistician, vol. 42, no. 1, pp. 59–66, 1988.

[6] R. K. Paul, "Multicollinearity: Causes, effects and remedies," IASRI, New Delhi, vol. 1, no. 1, pp. 58–65, 2006.

[7] E. R. Mansfield and B. P. Helms, "Detecting multicollinearity," The American Statistician, vol. 36, no. 3a, pp. 158–160, 1982.

[8] M. A. Schroeder, J. Lander and S. Levine-Silverman, "Diagnosing and dealing with multicollinearity," Western Journal of Nursing Research, vol. 12, no. 2, pp. 175–187, 1990.

[9] A. Amin, B. Shah, A. M. Khattak, T. Baker & S. Anwar, "Just-in-time customer churn prediction: With and without data transformation," IEEE Congress on Evolutionary Computation (CEC), pp. 1–6, 2018.

[10] R. Tibshirani, "Regression shrinkage and selection via the lasso," Journal of the Royal Statistical Society: Series B (Methodological), vol. 58, no. 1, pp. 267–288, 1996.

[11] G. James, D. Witten, T. Hastie and R. Tibshirani, "An introduction to statistical learning," New York: Springer, vol. 112, pp. 3-7, 2013.

[12] K. G. Pillay, S. S. S. M. F. Avtar and M. A. A. Abdullah, "Interaction Effects on Prediction of Children Weight at School Entry Using Model Averaging," International Journal of Engineering & Technology, vol. 7, no. 4.30, pp. 205, 2018.

[13] A. Schützenmeister, U. Jensen and H. P. Piepho, "Checking normality and homoscedasticity in the general linear model using diagnostic plots," Communications in Statistics-Simulation and Computation, vol. 41, no. 2, pp. 141–154, 2012.

[14] F. Changyong, W. Hongyue, L. U. Naiji, C. Tian, H. E. Hua and L. U. Ying, "Log-transformation and its implications for data analysis," Shanghai Archives of Psychiatry, vol. 26, no. 2, pp. 105, 2014.

[15] G. T. Knofczynski and D. Mundfrom, "Sample sizes when using multiple linear regression for prediction," Educational and Psychological Measurement, vol. 68, no. 3, pp. 431–442, 2008.