Middle-East Journal of Scientific Research 15 (4): 606-612, 2013
ISSN 1990-9233
© IDOSI Publications, 2013
DOI: 10.5829/idosi.mejsr.2013.15.4.11116

Statistical Approach for Modeling Malaysia's Gross Domestic Product

[1]*Ho Wei Chin and* [1,2,3]*Anwar Fitrianto*

[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Malaysia
[2]Laboratory of Applied and Computational Statistics,
Institute for Mathematical Research, Universiti Putra Malaysia
[3]Department of Statistics, Faculty of Mathematics and Natural Sciences,
Bogor Agricultural University (IPB), Indonesia

Abstract: This study investigates factors which affect Malaysia's gross domestic product. Stepwise regression has been used to construct the appropriate model. The model is constructed by splitting the data into model building set and validation set. The data splitting shows that the regression model is valid. After the process performing to determine the appropriate regression model, the variables private consumption, exports of goods and services and interest rate have significant contribution to Malaysia's GDP.

Key words:

INTRODUCTION

Gross domestic product (GDP) refers to the market value of all final goods and services produced within a country in a given period. GDP can be used as an indicator of a country's standard living. The downturn in worldwide trade in 2009 hit Malaysia particularly hard. GDP growth slowed down to 0.1% in the last quarter of 2008 and decelerated by -6.2% and -3.9% respectively in the first two quarters of 2009 as a consequence [1]. However, a robust recovery had done quickly. Momentum behind domestic private consumption and investment were being build building as the recovery in external demand continues. As a result, growth of 5.7% was projected for 2010, following a contraction of 1.7% in 2009. In this study, an appropriate multiple regression model will be employed to predict for Malaysia's GDP in the future. The model will be developed by using stepwise regression method. The objectives of the study are to examine the relationship between the independent variables with the dependent variables, gross domestic product (GDP), to determine independent variables that can affect on Malaysia's gross domestic product (GDP) and to derive an appropriate regression model that provides the most accurate forecasts for Malaysia's gross domestic product (GDP).

Gross Domestic Product (GDP): Gross Domestic Product (GDP) is a measure of a country's overall economic output. It is the market value of all final goods and services made within the borders of a country in a year. The pioneer of the GDP concept is Simon Kuznets in 1934, when he reported to the US Congress acknowledged the GDP's flaws as an economic indicator, "The welfare of a nation can scarcely be inferred from a measurement of national income", [2]. GDP can be determined in three ways, all of which should, in principle, give the same result. They are the product (or output) approach, the income approach and the expenditure approach. Product approach means how many goods and services were sold, income approach means how much income (profit) was earned and expenditure approach means how much money was spent.

To understand how the economy is using its scare resources, economists are often interested in studying the composition of GDP among various types of spending. In order to do this, GDP is divided into four components as personal consumption (C), gross investment (I), government spending (G) and net exports (NX). So, GDP can measure through the following famous equation:

$$GDP = C + I + G + NX.$$

Corresponding Author: Ho Wei Chin, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Malaysia

In this case, because each ringgit of expenditure included in GDP is placed into one of the four components of GDP, the total of the four components must be equal to GDP. Personal consumption (C) is the expenditure of households on different items such as food and medical expenses, for example, the Smith's lunch at Burger King. Gross investment (G) is the purchases of machinery, building new houses, buildings but it does not refer to investing in shares or bonds. Meanwhile, related to investment, [3] studied Influence of economic factors on performance of investment and mudharabah. Government spending (G) includes money spent on paying government employees, also include government's investment expenditure such as on roads. However, it does not include transfer payments such as retrials and subsidies. Net exports (NX) refer to the difference between the value of all exports and the value of all imports. Subtraction between values of exports and imports is made because imports of goods and services are included in other components of GDP.

In other words, net exports include goods and services produced abroad (with a minus sign) because these goods and services are included consumptions, investments and government purchases (with plus sign). Thus, when a domestic household, firm, or government buys a good or services from abroad, the purchases reduces net export but it also raises consumptions, investment or government purchases, it does not affect GDP. If exports exceed imports, it adds to the GDP. If not, it subtracts from the GDP. Thus, even if a nation's people work very hard to produce products for exports, but still import more than they export, the nation's GDP will be negatively impacted.

With regard to Malaysia's GDP, [4] pointed out that for Malaysia, there is an impact of GDP on Foreign Direct Investment (FDI) in the short run and positive connection in the long run. Another article about Malaysian's GDP is available in [5]. They found that Gross Domestic Product of manufacturing, trade openness, domestic credit to private sector and domestic direct investment significantly influenced the level of foreign direct investment inflow into Malaysia.

Multiple Linear Regression Model: According to [6], multiple regression is a flexible method of data analysis that may be appropriate whenever a quantitative variable (the dependent or criterion variable) is to be examined in relationship to any other factors (expressed as independent or predictor variables). Currently, there are many methods to conduct the multiple linear regression, but the most used method based on Least Squares (LS). One of the latest methods is artificial neural network (ANN). Pao [7] studied a comparison of neural network and multiple regression analysis in modelling capital structure. Zaefizadeh et. al. [8] compared between multiple linear regression and Artificial Neural Network for MLR and ANN methods to predict yield in barley.

Relationships may be nonlinear, independent variables may be quantitative or qualitative and one can examine the effects of a single variable or multiple variables with or without the effects of other variables taken into account.

The general form of multiple regression model is:

$$y_i = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \varepsilon_i \qquad (1)$$

where:

y_i is the dependent variable,
$x_1, x_2, ..., x_n$ are the independent variable,
$\beta_0, \beta_1, ..., \beta_n$ are the parameters,
ε_i is the error term.

In the model, it is assumed that the error terms are independent and normally distributed with mean 0 and variance σ^2.

Levine *et al.* [9] pointed out that multiple regression analysis mainly used to develop a statistical model that can be used to predict the values of a dependent or response variable based on the values of at least one independent variable. It is used to test the hypothesis of the relationship between dependent variable, *Y* with two or more independent variables, *X* and for prediction.

Multicollinearity Issues in Linear Regression: According to [10], multicollinearity is a state of very high intercorrelations or inter-associations among independent variables. Multicollinearity caused by several reasons, such as there has an inaccurate use of dummy variables, the inclusion of an almost identical variable twice, the inclusion of a variable which is computed from other variables in the equation, or when the variables are highly and truly correlated to each other. Several recent articles discussed about multicollinearity. Comparisons between three shrinkage regression models, Ridge Regression (RR), Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) were discussed in [11],

[12]. Irfan *et al* [12] found that those three shrinkage regression methods provide more informative results as compared to the Ordinary Least Square (OLS) method to handle the problem of multicollinearity on real GDP data in Pakistan and [13] used latent variable regression method to remedy the multicollinearity problem.

Pardoe [14] said that to solve the problem of multicollinearity, we can attempt to collect new data with a lower correlation between the collinear predictors; this is rarely going to be possible with observational data. Or if possible, combine the collinear predictors together to form a new predictor; this usually only makes sense if the predictors in question have the same measurement scale. We also can settle this problem by remove one of the collinear predictors from the model, for example, remove the one with the highest *p*[W76]-value, we necessarily lose some information in the dataset with this approach, but it may be the only remedy if the previous two approaches are not feasible.

In order to measure collinearity between independent variables, [10] suggested to use Variance Inflation Factor (VIF), which has the following equation:

$$VIF_j = \frac{1}{1-R_j^2}, \quad j = 1,2,...,p \tag{2}$$

where R_j^2 is the coefficient of the multiple determination of the independent variable $X_j$with all the others Xvariables. R_j^2 measures the proportion of the total variation in Y which is explained by the predictive power of all the independent variables through the multiple regression model. The computational formula for R_j^2 is

$$R_j^2 = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{3}$$

where *SST* known as total sum of square and *SSE* known as sum of squared error.

Many articles discussed about mullticollinearity and VIF. [15] pointed out that VIF tells us how "inflated" the variance of the coefficient is, compared to what it would be if the variable were uncorrelated with any other variable in the model. Meanwhile, [16] have shown how multicollinearity has increased the instability of the coefficient estimates. Regarding the cut-off point to indicate the presence of multicollinearity, it is recommended that a VIF greater than 10 indicates the presence of strong multicollinearity [17]. In the intervening time, [9] studied that if a set of independent variables is uncorrelated, then the VIF_j is equal to 1. If the set is highly correlated, then VIF_j might even greater than 10. Meanwhile, [18] suggested that all VIF_j values should be less than 5.

Model Validation: Model validation is the final step in the model building process. In this step, usually it involves checking the selected regression model against independent data. There have three ways on validating the regression model:

- Collection of new data set to check the model and its predictive ability.
- Comparison of results with theoretical expectations, empirical results and simulation results.
- Use hold-out sample to check the model and its predictive ability.

Collection of new data can be use to examine whether the regression model developed from the earlier data is still applicable for the new data. According to [10], means squared prediction error *(MSPR)* can be used to measure the performance of the regression model. The *MSPR* is calculated as follow:

$$MSPR = \frac{\sum_{i=1}^{n*}(Y_i - \hat{Y}_i)^2}{n*} \tag{4}$$

where Y_i is the value of the response variable in the i^{th}validation case, $\hat{Y}_i$ is the predicted value for the i^{th} validation case based on the modelbuilding data set and $n*$ is the number of cases in the validation data set.

Neter, *et al*. [10] also stated that when a data set is large enough to split data into two set, data splitting is an alternative to validate a regression. Data splitting is an effective way to simulates partial or complete replication of the study. When the data set splitting into two set, the first set is known as model-building model which is used to developed the model. While the second set is known as validation or prediction set which is used to evaluate the reasonableness and predictive ability of the selected model.

Data sets are not always split equally into the model-building and validation data set. It is important to take note that the model building data set should be sufficiently large in order to develop a reliable model. The number of observation in model-building set should be at least 6 to 10 times the number of variables in the pool of independent variables. If the entire data set not large enough to make an equal split, the validation data set should be smaller than the model-building set. The variance of the estimated regression coefficients developed from model-building set usually larger than those that obtained from the fit to the entire data set. Once the regression model has been validated, the entire data set can used to estimate the final regression model.

MATERIALS AND METHODS

Data: The data used for this study is obtained from the [19]. The data selected is from year 1980-2010. The data is made up by 31 observations of annual percentage of Malaysia's GDP, annual percentage of private consumption, annual percentage of government consumption, inflation rate and unemployment rate, annual percentage of imports and exports and interest rate.

There is only one continuous response variable and several independent variables in the study. The response variable, y, is the GDP. Meanwhile, several potential independent variables are private consumption, x_1 government consumption, x_2 inflation rate, x_3 unemployment rate, x_4 imports of goods and services, x_5 exports of goods and services, x_6 and interest rate, x_7.

Methodology: In this study, we use a well-known software-Minitab *Release* 16 to get the result of the data. This software is useful in stacking and sorting the data, creating patterned and random data, statistic analysis, graphical techniques, control charts and etc. Based on the result getting from this software, we start analyse the data.

To investigate the first objective in our study, first, we need to determine the correlation between the gross domestic products y with the independent variables and also the correlation between the independent variables. If any of the correlations between each pair of independent variables is greater than the highest correlations between gross domestic product and each of the independent variables, there is a potential that multicollinearity problem exists.

We also calculate variance inflation factors (VIF) for the regression parameter estimates for each of the independent variables in the model to check whether it exist the problem of multicollinearity. There have several rules of thumb associated with VIF. If the value of VIF is greater than 10, then the problem collinearity is consider severe. When the VIF reaches its threshold value of 10, we can reduce the collinearity by eliminating one or more independent variables into single index.

After eliminating the highly correlated variables from the regression model, we can proceed to the stepwise regression. Through stepwise regression, we can obtain a suitable multiple regression models. Then, for achieving the second objective in this stud, we need to observe the results of regression analysis in the previous step to identify the independent variables which have significant influence on gross domestic product. In summary, the steps of the analysis are as follows:

Step 1: Conduct correlation analysis using complete data to identify linear correlations between independent variables and between independent and dependent variables,

Step 2: Conduct regression analysis using the complete datas*et al*ong with obtaining VIF values to identify possible highly correlated independent variables,

Step 3: Do the stepwise regression analysis using the complete dataset to identify independent variables which has significant contribution to the response variable,

Step 4: Validate the dataset into two sets, namely model building and validation dataset,

Step 5: Analyse and compare both model building and validation dataset.

RESULTS AND DISCUSSION

Relationship Between The Independent And Dependent Variables: In order to determine the relationship between the independent variables and dependent variable, we need to construct a correlation matrix. In this study, it consists of seven independent variables, $x_1, x_2, x_3, x_4, x_5, x_6, x_7$ and a dependent variable, y. So, there are $C_2^8 = 28$ zero-order a correlation, the following is the correlation matrix getting from this study:

From Table 1, the correlation matrix shows that there are correlations between each of the independent variables and between independent and dependent variable. For example, there is a strong linear relationship between aount of imports of goods and services, x_5 and the response variable (r=0.841). Meanwhile, linear correlation between inflation rate, x_3 and amount of imports of goods and services, x_5, is very low (r = 0.032) as we expect to have for correlations between independent variables.

In order to proceed to the next step, it is very important that the independent variables show lower correlated between each others. This means that there is no redundant variable using in the model. This can prevent the redundant variable exist and affect the result of the model. We need to make sure the highly correlated variables have been removed from the model. Meanwhile, those independent variables which appear to correlate well with y, the outcome variable are to be the more likely candidates for inclusion.

Statistical Model Building Using Stepwise Regression Model: Before continuing to the stepwise regression, we need to confirm that there is no highly correlated independent variables exist in the model. So, we perform a regression analysis which is displayed in Table 2 to check its Variance Inflation Factor (VIF). In this study, since there is more than one term in the model, we observe R-squared adjusted value instead of R-squared value. From Table 2, we can see that the value of R-squared adjusted is 78.2%. This tells us that 78.2% of the variation in gross domestic product, x_5 is explained by the regression model. This figure also show the VIF value of imports of goods and services, x_5 is 22.781 which is larger than 10. This means that the imports of goods and services, x_5 are highly correlated with the others independent variables. There is a problem of collinearity of x_5 and the other variables. So, we can try to solve this problem by removing imports of goods and services, x_5, from the model. Then, we can proceed to the stepwise regression. Analyzing using stepwise procedures give us results as displayed in Table 3.

Table 3 shows the stepwise regression on this model. At the first step of stepwise regression, private consumption, x_1 is entered to the model with $\hat{\beta}_0 = 2.3538$ and $\hat{\beta}_1 = 0.623$, to obtain estimated regression of the following equation, which has R-squared of 67.54% :

$$\hat{y} = 2.3538 + 0.623x_1 \cdot$$

Table 1: Correlation Matrix and significance of all variables in the study

	y	x_1	x_2	x_3	x_4	x_5	x_6
x_1	0.784 (0.000)						
x_2	0.322 (0.095)	0.383 (0.044)					
x_3	0.100 (0.594)	0.013 (0.945)	0.221 (0.259)				
x_4	-0.156 (0.437)	-0.260 (0.190)	-0.262 (0.187)	-0.247 (0.215)			
x_5	0.841 (0.000)	0.821 (0.000)	0.360 (0.060)	0.032 (0.873)	-0.087 (0.665)		
x_6	0.608 (0.001)	0.371 (0.052)	-0.172 (0.383)	-0.204 (0.297)	0.087 (0.666)	0.725 (0.000)	
x_7	0.105 (0.643)	-0.278 (0.211)	0.377 (0.092)	-0.064 (0.777)	0.074 (0.750)	-0.137 (0.553)	-0.149 (0.519)

Note: values in bracket are p values for respective correlation coefficient

Table 2: Estimated regression coefficients of the initial steps of model building

Predictor	Coefficient	SE Coefficient	t	p	VIF
Constant	0.372	3.208	0.12	0.91	
x_1	0.6081	0.2135	2.85	0.014	7.286
x_2	-0.06	0.1264	-0.47	0.643	2.755
x_3	-0.3464	0.517	-0.67	0.515	2.756
x_4	-0.3928	0.3799	-1.03	0.32	2.266
x_5	-0.0386	0.1677	-0.23	0.822	22.781
x_6	0.262	0.1907	1.37	0.193	9.621
x_7	0.6113	0.2304	2.65	0.02	2.112

Note: $R^2_{adj} = 0.782$

Table 3: Model selection steps using stepwise regression procedures

	Terms in the model						
Steps	Con-stant	x_1	x_7	x_6	s	R^2	R^2_{adj}
1	2.3538	0.623			2.33	67.54	65.83
2	-0.2898	0.695 (p=0.000)	0.49 (p=0.014)	0.189 (p=0.013)	2.02	77.05	74.5
3	-1.3011	0.559 (p=0.000)	0.48 (p=0.006)	0.189 (p=0.013)	1.72	84.23	81.45

Note: Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

For second step, the interest rate, x_7 is also entered to the model with parameters $\hat{\beta}_0 = -0.2898, \hat{\beta}_1 = 0.695$ and $\hat{\beta}_7 = 0.49$. The R-squared adjusted is increased to 74.50%. In the final step of thestepwise regression, the exports of goods and services, x_6 is entered the model with $\hat{\beta}_0 = -1.3011, \hat{\beta}_1 = 0.559, \hat{\beta}_7 = 0.48$ and $\hat{\beta}_6 = 0.189$. The R-squared adjusted also increased to 81.45%. The model is become as follows:

$$\hat{y} = -1.3011 + 0.559x_1 + 0.48x_7 + 0.189x_6 \cdot$$

Table 4: Comparison between Model Building Set and Validation Set

Statistics	Model Building Set	Validation Set
$\hat{\beta}_0\left(SE\left(\hat{\beta}_0\right)\right)$	0.053 (1.946)	-0.7678 (0.8939)
$SE\left(\hat{\beta}_0\right)$	1.946	0.8939
$\hat{\beta}_1\left(SE\left(\hat{\beta}_1\right)\right)$	0.3768 (0.1623)	1.03927 (0.09741)
$\hat{\beta}_6\left(SE\left(\hat{\beta}_6\right)\right)$	0.24483 (0.06937)	-0.4446 (0.1212)
$\hat{\beta}_7\left(SE\left(\hat{\beta}_7\right)\right)$	0.3638 (0.1958)	1.2523 (0.2079)
SSE	21.388	3.110
PRESS	62.6026	-
MSE	2.376	0.778
MSPR	-	2.7759
R^2_{adj}	66.1%	97.6%

After the several stages on determine the suitable variables of the model, we come out a regression model with three independent variables which is x_1,x_6 and x_7 as the above regression model.

Validating the Regression Model: In order to validate the regression model, we separate the data set to two set which is model-building set and validation set. A comparison table between model-building and validation set is displayed on as Table 4 in order to clearly stated out the data between this two set. From the Table 4, we can see that for the model building data set *MSE* = 2.376 is fairly close to *MSPR* = 2.7759 in the validation set. According to Neter *et al* (1996), when the *MSE* in model building data set is close to *MSPR*, this means that the following obtained model:

$$y = -1.3011 + 0.559x_1 + 0.189x_6 + 0.48x_7$$

is suitable for prediction for the future GDP.

CONCLUSION

The purpose of this study is to build a multiple regression model to forecast Malaysia's gross domestic product in the future. This study investigates the internal factors that can affect Malaysia's gross domestic product. After several steps in model building and model validation, we obtain the following final regression model obtained:

$$y = -1.3011 + 0.559x_1 + 0.189x_6 + 0.48x_7$$

There are three variables that have relationship with GDP, that is private consumption, x_1, exports of goods and services, x_6 and interest rate, x_7. If increase of 1 unit in private consumption, x_6, this make the GDP increase by 0.559; if increase of 1 unit in exports of goods and services, x_6, this will make the GDP increase by 0.189; similarly, if increase of 1 unit in interest rate, x_7, the GDP will increase by 0.48.

REFERENCES

1. Abidin, M.Z. and R. Rajah, 2009. The Global Financial Crisis and the Malaysian Economy: Impact and Responses. United Nations Development Programme (UNDP), Malaysia.
2. Costanza, R., H. Maureen, P. Stephen and T. John, 2009. Beyond GDP: The Need for New Measures of Progress. Working Papers. Boston University, Pardee House, Boston.
3. Zainal, N.S. and Z.M. Yusof, 2009. Influence of Economic Factors on Performance of Investment and Mudharabah Accounts in Maybank, Malaysia, International Journal of Economics and Finance. 1(2): 221-224.
4. Fizari, F., H.A. Abu, S. Worazidah, A.H.A.K. Rhaudhah, S.B. Nurul, A. Salwani and J. Kamaruzaman, 2011. Impact of Export and Gross Domestic Product Towards Foreign Direct Investment Inflows in Malaysia. World Applied Sciences Journal, 12(Special Issue on Bolstering Economic Sustaimbility): 27 -33.
5. Nezakati, H., F. Farzad and M.V. Behzad, 2011. Do Local Banks Credits to Private Sector and Domestic Direct Investments Affect FDI Inflow? (Malaysia Evidence).World Applied Sciences Journal, 15(11): 1576-1583.
6. Pearson, K. and A. Lee, 1908. On the Generalized Probable Error In Multiple Normal Correlation, Oxford Journals, 6(1): 59-68.
7. Pao, H.T., 2008. A comparison of neural network and multiple regression analysis in modelling capital structure, Expert Systems with Applications, 35: 720-727.
8. Zaefizadeh, M., K. Majid and G. Roza, 2011. Comparison of Multiple Linear Regressions (MLR) and Artificial Neural Network (ANN) in Predicting the Yield Using its Components in the Hulless Barley. American-Eurasian J. Agric. and Environ. Sci., 10(1): 60-64.

9. Levine, D.M., T.C. Krehbiel and M.L. Berenson, 2003. Business Statistics: A First Course. (3^{rd} Eds). Prentice Hall, New Jersey.
10. Neter, J., M.H. Kutner, C.J. Nachtsheim and W. Wasserman, 1996. Applied Linear Regression Models. (3^{rd} Eds). Times Mirror Higher Education, Toronto.
11. Irfan, M., J. Maria and A.R. Muhammad, 2013. Comparison of Shrinkage Regression Methods for Remedy of Multicollinearity Problem. Middle-East Journal of Scientific Research, 14(4): 570-579.
12. Ramzan, S.F., M.Z. Aisal and R. Shumila, 2010. Prediction Methods for Time Series Regression Models with Mnlticollinearity. World Applied Sciences Jownal, 11(4): 443-450.
13. Ramzan, S. and I.K. Muhammad, 2010. Dimension Reduction and Remedy of Multicollinearity using Latent Variable Regression Methods. World Applied Sciences Jowna, 18(4): 404-410.
14. Pardoe, I., 2006. Applied Regression Modeling: A Business Approach. John Wiley and Sons, New Jersey.
15. Allison, P.D., 1999. Multiple Regression – A Primer. Pine Forge Press, Thousand Oaks, CA.
16. Freund, Rudolf, J. and R.C. Littell, 2000. SAS System for Regression. (3^{rd} Eds). SAS Institute, Cary, NC.
17. Marquardt, D.W., 1980. The Importance of Statisticians. Journal of the American Statistical Association. 75: 87-91.
18. Snee, R.D., 1973. Some Aspects of Nonorthogonal Data Analysis, Part I. Developing Prediction Equation, Journal of Quality Technology. 5: 67-69.
19. Bank Negara Malaysia, 2011.Economic and Financial Data for Malaysia. http://www.bnm. gov.my/index. php?ch=statistic_nsdpanduc=2 (accessed on 12 February 2010)