Labuan Bulletin of International Business &Finance

Volume 16 Issue 1 eISSN 2600- 7894

ESTIMATION OF TRADE BALANCE USING MULTIPLE LINEAR REGRESSION MODEL

Sarimah Omar Gan[a1], Sabri Ahmad[b]

[a]*Labuan Faculty of International Finance, Universiti Malaysia Sabah, Labuan International Campus, Jalan Sungai Pagar, 87000 Labuan F.T., Malaysia*
[b]*School of Informatics and Applied Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Malaysia*

ABSTRACT

This study aims to evaluate the performance of multiple linear regression in estimating trade balance, so that a regression model for estimating the trade balance can be developed based on the important variables that have been identified. The performance of four regression methods including enter, stepwise regression, backward deletion, and forward selection is measured by mean absolute error, standard deviation, and Pearson correlation at the validation stage. The study concludes that multiple linear regression model developed by stepwise method is the best model for the trade balance estimation. The model considers the following six significant variables: Exports of palm oil, imports of tubes, pipes, and fittings of iron or steel, exports of crude petroleum, imports of petroleum products, exports of plywood plain, and imports of rice. The regression model achieves a moderate value of model estimated accuracy (76.10%), mean absolute error (0.257), standard deviation (0.308), and linear correlation (0.851).

***JEL classification**: C02.*
***Keywords**: Export and import totals; external trade; multiple linear regression model; trade balance.*

1. INTRODUCTION

Sabah is the second largest state in Malaysia after Sarawak and it has a total land area of 73,610 square kilometres. The state is located to the northern region of Borneo Island, and it is strategically situated between the wealthy markets of North Asia and the fast developing regions of Southern Philippines, Brunei and Kalimantan. Sabah could exploit the market liberalisation driven by the ASEAN Free Trade Area (AFTA) for transhipment and add value to cargo between the developing countries in Southeast Asia and wealthy North Asia. The state is also rich in natural resources including oil, natural gas, minerals, forestry, fertile agriculture land and marine life. The main resource of the state's economy was timber export but palm oil trade has emerged as an alternative trade in line with the efforts to promote forest conservation. The major exports including cocoa beans, crude oil petroleum, palm oil products, plywood, and sawn timber. The main imports are including fertilizers, motorcars,

[1]Corresponding author's email: sarimahgan@ums.edu.my

petroleum products, sugar, rice, and tubes or pipes. The state is heavily dependent on external trade where the trade balance stands a relatively high portion of its gross domestic product (GDP). Sabah's total external trade in 2016 was RM70.26 billion, comprising RM41.4 billion in exports and RM28.95 billion in imports (Department of Statistics Malaysia, 2017).

Trade balance is one of the key components of a country's GDP formula. GDP increases when the total value of goods and services produced in one country and purchased by citizens of another country exceeds the total value of foreign goods and services brought into one country from another country, otherwise known as a trade surplus. GDP decreases when the domestic consumers spend more on foreign products than domestic producers sell to foreign consumers, otherwise known as a trade deficit. To success of all businesses, it is important to put a planning for the future, and the technique is called forecasting. It includes estimating important factors that could influence outcomes for a business. Therefore, this study helps identify the important variables that give a significant contribution to the trade balance, and hence a forecasting model can be developed to estimate the trade balance.

2. METHODOLOGY

There are many variables contributing the balance of trade and we have to be selective in this study to ensure a significant model can be developed. In this study, we propose the development of forecasting model for determining balance of trade using multiple linear regression (MLR). MLR is examined for a range of monthly data of external trade in Sabah in term of import and export totals. The data used in this study are the secondary data of external trade in Sabah for the period of 2005 to 2014. Monthly data for ten years have been obtained from Institute for Development Studies (Sabah). There are three types of data including exports, imports and trade balance. The data are measured in local currency, the Ringgit Malaysia (RM). These data are analysed by using SPSS Statistics and SPSS Clementine. The performance of four regression methods including enter, stepwise, backwards, and forwards is measured by mean absolute error, standard deviation and Pearson correlation at the validation stage. The main objective of this study is to develop the best forecasting model for trade balance estimation by evaluating the performance of the four regression methods and making a comparison, and hence the significant variables in determining the trade balance can be identified.

There are a few measures of accuracy showed in the past forecasting literature. There can be many performance measures for the data mining forecaster such as the modeling time and training time, the ultimate and the most important measure of performance is the prediction accuracy it can achieve beyond the training data (Zhang et al., 1998). An accuracy measure can be defined as the difference between the actual (desired) and the predicted value (Zhang et al., 1998). In this study, we are using four types of accuracy measures to evaluate the performance of the regression model, namely, mean error (ME = $\frac{\sum(e_t)}{N}$), mean absolute error (MAE = $\frac{1}{N}\sum|e_t|$), mean squared error (MSE = $\frac{\sum(e_t)^2}{N}$), and root mean squared error (RMSE = $\sqrt{\frac{\sum(e_t)^2}{N}}$). Besides that, the standard deviation, Pearson correlation to the target variable, and estimated accuracy are also calculated as a measurement of the predictive performance of the regression model.

In this study, we have identified several variables which influence the balance of trade such as exports of cocoa beans, exports of crude petroleum, exports of palm oil, exports of plywood plain, exports of sawn timber, imports of fertilizers (manufactured), imports of motor cars (completely built-up), imports of petroleum products, imports of refined beet and cane sugar, imports of rice, and imports of tubes, pipes, and fittings of iron or steel. The dependent variable and eleven independent variables in this study are represented as:

One dependent variable:
i. b1 = Balance of trade

Eleven independent variables:
i. e1 = Exports of cocoa beans
ii. e2 = Exports of crude petroleum
iii. e3 = Exports of palm oil
iv. e4 = Exports of plywood plain
v. e5 = Exports of sawn timber
vi. i1 = Imports of fertilizers (manufactured)
vii. i2 = Imports of motor cars (completely built-up)
viii. i3 = Imports of petroleum products
ix. i4 = Imports of refined beet and cane sugar
x. i5 = Imports of rice
xi. i6 = Imports of tubes, pipes, and fittings of iron or steel

The raw data have undergone a process of mathematical transformation which is through an arithmetic operation, $\ln(x)$. Data transformation is made so that the data will be distributed more evenly. The external trade data are then partitioned into a training dataset and a test dataset. A training dataset is implemented to build up a model while a test dataset is to assess the performance of the model (Zhang et al., 1998). The division of the data into the training and test dataset is made based on the predetermined percentage of the ratio of 60% training to 40% test (case 1), 70% training to 30% test (case 2), 80% training to 20% test (case 3), and 90% training to 10% test (case 4). Next, the four cases are tested and compared. The case with a lower mean absolute error rate is chosen for further analysis. The selection of the training and test sample may affect the performance of the forecasting model (Zhang et al., 1998). The models are validated to ensure that they are adequate to be deployed for predictive purposes. The research methodology is summarised in Table 1.

Table 1: Research methodology.

Step	Description
Step 1	Split the data into training sample (60%, 70%, 80%, 90%) and test sample (40%, 30%, 20%, 10%), respectively.
Step 2	Test and compare the four cases. The case with a lower mean absolute error rate is chosen for further analysis.
Step 3	Build the MLR models (enter, stepwise, backwards, forwards) using training dataset.
Step 4	Evaluate the performance of each regression method using validation dataset (mean absolute error, standard deviation, Bivariate (Pearson) correlation). The method with a lower mean absolute error rate is chosen as the suitable method for the MLR model.

Table 1 continued.

Step	Description
Step 5	Measure the predictive performance of the chosen MLR model using validation dataset (ME, MAE, MSE, RMSE, standard deviation, Bivariate (Pearson) correlation, estimated accuracy).
Step 6	Develop the best regression model and then identify the significant variables.

2.1 Multiple linear regression model

MLR analysis is a statistical method that is widely used among researchers to associate variables with the data being analysed and the method is used to determine the linear relationship between a response variable and explanatory variables (Bowerman et al., 2005). Independent variables referred as the explanatory variables while a dependent variable referred as a response variable (Rencher & Schaalje, 2008). MLR examines how several explanatory variables are related to one response variable. Once the significant variables have been identified to forecast the response variable, the information on the multiple variables can be used to produce a high accuracy of forecast. MLR model oftenly used to examine some proposed theoretical model, the model creates a relationship in a form of straight line (linear) that best approximates all the individual data points (Zikmund et al., 2010). The use of this method has some limitations in which it can only be applied when there is a linear relationship between the response variable and the explanatory variables (Singh & Prajneshu, 2008).

Forecasting future in constant changes in the economy and capital market is one of the important issues which is an often discussed topic among financial researchers (Moradzadehfard et al., 2011). The classical method such as regression has contributed a relative success in these fields, the regression method has been considered because the estimated coefficients can be interpreted easily (Moradzadehfard et al., 2011). Economic performance is to be evaluated on the basis of effectiveness and efficiency of use of resources. Shyti et al. (2016) studied the economical phenomena through a statistical point of view and they emphasised the validity of the regression analysis in economic performance. They found that number of employees and price of product have a big impact on the level of sales revenue. The income on the product is the result of conjugation many influencing variables, but not all the determined ratios have the same importance, the action of some of them compensating each other. Sopipan (2013) using MSE and MAE as accuracy measures to evaluate and compare the performance of three forecasting returns for the Stock Exchange of Thailand (SET) Index. He found that multiple regressions based on principal component analysis has the best performance with the lowest MSE (0.8886) and MAE (0.7463).

In short, regression analysis describes the behavior of the dependent variable given the value of independent variables (Mokhtar, 1994). MLR model is a linear model that predicts the relationship between independent and dependent variables. There are four important steps in conducting the MLR model. The steps are checking assumptions, selecting a suitable method of MLR, interpreting the output, and developing the regression equation (Sarimah & Sabri, 2011). The evaluation on the aptness of a regression model must be performed before further analysis is made. It is very important to make sure the behaviour of the residuals has met the underlying assumptions for the error values in the model. The model is not appropriate for the data if any of the statistical assumptions of the model are not met (Matson &

Huguenard, 2007). In this study, the residual analysis has been performed and the four assumptions of a linear model such as the error terms are normally distributed, the dependent and independent variables have a linear relationship, the error terms have constant variance, and the error terms are independent have been obtained. Next, the four regression methods including enter, stepwise regression, backward deletion, and forward selection are used to build the MLR model. In stepwise multiple regression analysis, the researcher provides a list of explanatory variables and then allows the program to select which valiables it will enter and in which order they go into the equation, based on a set of statistical criteria (Pallant, 2013). The value of regression coefficients depends upon the variables in the model, therefore the selection of explanatory variables should be done properly, this means that the explanatory variables included in the model and the way in which they are entered into the model can have a great impact. Stepwise regression builds the equation in steps in which each variable is entered in sequence based on a purely mathematical criterion and its value assessed (Field, 2000). Backward deletion involves starting with all explanatory variables are entered, testing the elimination of each variable using a chosen model fit criterion, then the weakest explanatory variables are removed one by one, the process is repeating until no further variables can be deleted without a statistically significant loss of fit (Tabachnick & Fidell, 2013). In forward selection, the strongest explanatory variables are entered one by one, testing the addition of each variable using a chosen model fit criterion, the process is repeating until none improves the model to a statistically significant extent (Tabachnick & Fidell, 2013).

The general descriptive form of a multiple linear equation is shown in equation (1). We use j to represent the number of independent variables. The estimated accuracy of the model can be explained from the value of coefficient of multiple determination, R^2. Coefficient of multiple determination is defined as the percent of variation in the dependent variable explained by the variation in the set of independent variables. R^2 can be calculated from the information we found in the ANOVA table and the formula is shown in equation (2).

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_j X_j \quad (1)$$

where:

$\hat{Y}$ is the dependent variable.

β_0 is the intercept, the value of $\hat{Y}$ when all the $X's$ are zero.

β_j is the regression coefficient, the amount by which $\hat{Y}$ changes when that particular X_j increases by one unit, with the values of all other independent variables held constant.

X_j is the independent variables.

$$R^2 = \frac{SSR}{SS\ total} \quad (2)$$

where:

R^2 is the coefficient of multiple determination.

SSR is the regression sum of squares.

$SS\ total$ is the total sum of squares.

3. RESULTS AND DISCUSSION

The process flow diagram for predictive modeling of MLR is shown in Figure 1. The MAE for the sample data partition is summarised in Table 2. Case 2 shows a more favorable performance compared to other cases as it produces the lowest MAE for all the four regression methods. In case 2, 84 out of 120 total number of external trade data are used as training sample to build the forecasting model, while the remaining data of 36 are used as a test sample to assess the performance of the model. Therefore, the ratio of 70% training and 30% is chosen as the most suitable partition to analyse the data of this study.

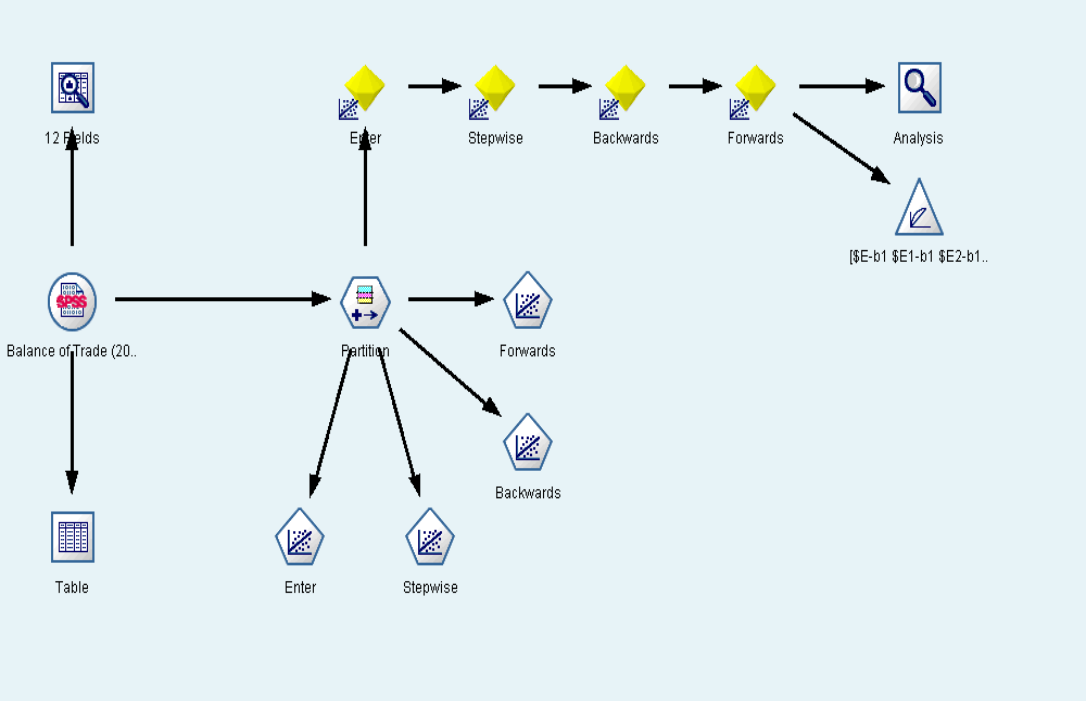

Figure 1: Process flow diagram of MLR.

Table 2: MAE of data partition.

Method	Case 1	Case 2	Case 3	Case 4
Enter	0.321	0.284	0.320	0.353
Stepwise	0.299	0.257	0.262	0.257
Backwards	0.315	0.279	0.311	0.342
Forwards	0.299	0.257	0.262	0.257

Next, by using the validation dataset through the selected data partition, a comparison between the four regression methods is performed based on the value of MAE, standard deviation and Bivariate (Pearson) correlation. The most suitable regression method will be chosen to build the forecasting model. The MLR models accuracy is summarised in Table 3. From the table summary, it shows model 2 shares the same result with model 4. Both models produce the best results with the lowest MAE (0.257). A low value of standard deviation (0.308) indicates that the regression method can build a regression model with a high level of predictive accuracy. A high value in linear correlation (0.851) indicates that there is a strong positive relationship

between the independent variables and the balance of trade. The two models show that there are six important variables that contribute significantly to the regression model such as exports of palm oil, imports of tubes, pipes, and fittings of iron or steel, exports of crude petroleum, imports of petroleum products, exports of plywood plain, and imports of rice. Both model 2 and model 4 show that there are 76.10% of variation in the prediction of trade balance explained by the variation in the set of six independent variables. Therefore, based on the rule of parsimony, model 2 is chosen as the most suitable regression model to estimate the balance of trade. The variable will be retained in the model if it contributes to the model, but all other variables in the model are then retested to see if they are still contributing to the success of the model, the variables will be removed if they are no longer contributing significantly (Chua, 2009). Finally, the values of regression coefficient in model 2 will be taken from the coefficient table in Table 4 in developing a regression equation. The chosen model can be defined as at the following regression equation shown in equation (3).

$$\hat{Y} = -15.649 + 1.413lnX_1 - 0.325lnX_2 + 0.827lnX_3 - 0.664lnX_4 + 0.395lnX_5 - 0.111lnX_6 \quad (3)$$

where:

$\hat{Y}$ is the trade balance.
X_1 is the exports of palm oil.
X_2 is the imports of tubes, pipes, and fittings of iron or steel.
X_3 is the exports of crude petroleum.
X_4 is the imports of petroleum products.
X_5 is the exports of exports of plywood plain.
X_6 is the imports of rice.

Table 3: MLR models summary.

Model	Method	MAE	Standard Deviation	Bivariate (Pearson) correlation	R^2	Significant Variables
1	Enter	0.284	0.376	0.817	0.727	All
2	Stepwise	0.257	0.308	0.851	0.761	e3, i6, e2, i3, e4, i5
3	Backwards	0.279	0.369	0.810	0.720	e2, e3, e4, e5, i1, i2, i3, i6
4	Forwards	0.257	0.308	0.851	0.761	e3, i6, e2, i3, e4, i5

Table 4: Coefficients[a].

Model		Unstandardised Coefficients		Sig.
		β	Std. Error	
2	(Constant)	-15.649	2.330	0.000
	Exports of palm oil	1.413	0.171	0.000
	Imports of tubes, pipes, and fittings of iron or steel	-0.325	0.093	0.001
	Exports of crude petroleum	0.827	0.117	0.000
	Imports of petroleum products	-0.664	0.112	0.000
	Exports of exports of plywood plain	0.395	0.152	0.011
	Imports of rice	-0.111	0.054	0.040

[a.] Dependent Variable: Balance of trade

4. CONCLUSION

Financial forecasters use numerous methods to arrive at their estimates. In this study, we can estimate the trade balance by developing a forecasting model using MLR. Initially, there are eleven potential independent variables that have been identified to estimate the trade balance. In the final analysis, the study concludes model 2 is the best forecasting model to estimate the trade balance. The model is developed by using stepwise regression method. In term of prediction, model 2 achieves moderate level of predictive accuracy. Model 2 shows only six independent variables contribute significantly to the model. The chosen regression model seems to be useful for estimating the trade balance with 76.10% of the variance in the dependent variable (trade balance) which can be predicted from the six independent variables (exports of palm oil, imports of tubes, pipes, and fittings of iron or steel, exports of crude petroleum, imports of petroleum products, exports of plywood plain, and imports of rice). The six important variables are considered the strengths in the external trade activity. To ensure these strengths are fully harnessed, the study recommend a competitive package of incentives should be provided to attract private sector investment in the promoted sectors. Besides that, emphasis should be placed on human capital development to support the growth of the targeted sectors. This study utilised the MLR model and we know that there are limitations in analysing the nonlinearity of economic and financial data by using the regression techniques. It is not easy to find data that meet all the assumptions of the regression model. Therefore, a more robust model and other soft computing techniques are recommended in future research.

REFERENCES

Bowerman, B. L., O'Connell, R. T. & Koehler, A. B. (2005). *Forecasting, time series, and regression: an applied approach* (4th ed.). California: Thomson Brooks/Cole.

Chua, Y. P. (2009). *Statistik penyelidikan lanjutan II: Ujian regresi, analisis faktor dan analisis SEM*. Kuala Lumpur: McGraw-Hill.

Department of Statistics Malaysia. (2014). Monthly Statistical Bulletin Malaysia. Malaysia.

Field, A. (2000). *Discovering statistics using SPSS for windows: Advanced techniques for the beginner*. Thousand Oaks: Sage Publications.

Matson, J. E. & Huguenard, B. R. (2007). Evaluating aptness of a regression model. *Journal of Statistics Education*, 15(2), 1-15.

Mokhtar, A. (1994). *Analisis regresi*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Moradzadehfard, M., Motlagh, P. A. & Fathi, M. R. (2011). Comparing neural network and multiple regression models to estimate dividend payout ratio. *Middle-East Journal of Scientific Research, 10*(3), 302-309.

Pallant, J. (2013). *A step by step guide to data analysis using IBM SPSS* (5th ed.). New York: McGraw-Hill.

Rencher, A. C. & Schaalje, G. B. (2008). *Linear models in statistics* (2nd ed.). Hoboken, New Jersey: John Wiley.

Sarimah, O. G. & Sabri, A. (2011). Multiple linear regression to forecast balance of trade. *Journal of Fundamental Sciences, 7*(2), 150-155.

Shyti, B., Isa, I. & Paralloi, S. (2016). Multiple regressions for the financial analysis of Alabanian economy. *Academic Journal of Interdisciplinary Studies, 5*(3), 300-304.

Singh, R. K. & Prajneshu. (2008). Artificial neural network methodology for modelling and forecasting maize crop yield. *Agricultural Economics Research Review*, 21, 5-10.

Sopipan, N. (2013). Forecasting the financial returns for using multiple regression based on principal component analysis. *Journal of Mathematics and Statistics*, *9*(1), 65-71.

Tabachnick, B. G. & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson.

Zhang, G. Q., Patuwo, B. E. & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, *14*, 35-62.

Zikmund, W. G., Babin, B. J., Carr, J. C. & Griffin, M. (2010). *Business research methods* (8th ed.). Canada: Cengage Learning.