Multiple Linear Regression to Forecast Balance of Trade

Sarimah Omar Gan[*] and Sabri Ahmad

Department of Mathematics, Faculty of Science and Technology, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia

ABSTRACT

The main objective of this study is to build a regression model by using multiple linear regression (MLR) analysis. MLR will be used when there are two or more controlled variables involved in the relationship. There are four general steps in building a regression model which are checking assumptions, selecting suitable methods of MLR, interpreting the output and selecting the best MLR model. The objective will be evaluated by using a time series data that have been obtained from Monthly Statistical Bulletin Sabah, Department of Statistics Malaysia, Sabah which is from year 2003 to 2009. The monthly data of external trade in Sabah in term of import and export totals for seven years will be analysed by using Statistical Package for the Social Sciences (SPSS) Statistics Version 17.0. Balance of trade for year 2011 and above can be forecasted through the regression model that has been developed by using MLR analysis.

| Balance of trade | import and export totals | multiple linear regression | regression model |

1. INTRODUCTION

Sabah is the second largest State in Malaysia with a total land area of 73,610 sq. km. Sabah is rich with natural resources from forest, mineral, flora, fauna to marine life. Forest resources and agriculture produce have always been the main sources of income for the State. Sabah has a very open economy in that it is relying on external trade too heavily. This is reflected in the high percentages of export (80.8%) and import (61.7%) to its GDP. Sabah's total external trade increased from RM21.2 billion in 1996 to RM27.3 billion in 2002 [1]. The economy of Sabah has always been heavily dependent on the export of its primary and minimally processed commodities. These major export items include palm oil products, cocoa beans, sawn timber, plywood and crude oil petroleum. The main imports of which include sugar, rice, petroleum products, tubes or pipes, motorcars, and fertilizers.

As mentioned above, Sabah has a very open economy that it is heavily dependent on external trade. In order for the State to handle the external trade activity more efficiently, they need to have enough income to cover all the expenses. Hence, this study may help the State especially the external trade department to analyse the strength of the trade activity and get some ideas so that they can see the opportunity of getting more income by increasing the exports activity.

This study will also help the department to be more cautious as well as to produce a good decision to manage their import-export activities during recession. The information that will be discovered in this study will be very useful for the State in their day to day trade activity.

2. MATERIALS AND METHODS

2.1 Data

A secondary data from year 2003 to 2009 have been used in this study. The monthly data of external trade in Sabah in term of import and export totals for seven years have been obtained from Monthly Statistical Bulletin Sabah, Department of Statistics Malaysia, Sabah. The data will be analysed by using Statistical Package for the Social Sciences (SPSS) Statistics Version 17.0.

2.2 Research method

The general objective of regression analysis is to set up a useful relationship between a dependent variable y and one or more predictor variables. Estimating a mean y value, forecasting an individual y value, and gaining an understanding into how changes in independent variable values affect y [2].

There are three main types of multiple regression analyses which are the standard multiple regression, hierarchical multiple regression and stepwise multiple regression. The most commonly used multiple regression analysis is the standard multiple regression where all the predictor variables are entered into the equation

Corresponding author at:
Department of Mathematics, Faculty of Science and Technology,
Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia
E-mail addresses: sarimahgan@hotmail.com (Sarimah Gan)

simultaneously. While in the hierarchical or sequential regression, the predictor variables are entered into the equation in the order specified by the researcher based on theoretical grounds. The stepwise multiple regression is a type of multiple regression where the researcher provides SPSS with a list of predictor variables and then allows the program to select which variables it will enter, and in which order they go into the equation, based on a purely mathematical criterion. There are three different versions of this approach which are the forward selection, backward deletion and stepwise regression [3]. In this study, the stepwise multiple regression method has been chosen to apply in the data analysis with the stepwise regression approach. Balance of trade will be the criterion variable. On the other hand, the eleven predictor variables are 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. There are four general steps in building forecasting model of trade balance in MLR. The general steps are checking assumptions, selecting suitable methods of MLR, interpreting the output and developing equation of MLR.

Linear regression models are tied to certain assumptions about the distribution of error terms. The model is not useful for making inferences if those assumptions are seriously violated. Therefore, before undergo further analysis based on the model is undertaken, it is important to consider the aptness of the model. Model aptness refers to the conformity of the behaviour of the residuals to the underlying assumptions for the error values in the model. The effective way of checking the assumptions is by residual analysis. Firstly, the error terms are normally distributed. Secondly, the dependent and independent variables have a linear relationship. Thirdly, the error terms have constant variance. Fourthly, the error terms are independent [4].

2.3 Step 1: Checking assumptions

When a regression model is built from a set of data, it must be shown that the model meets the statistical assumptions of a linear model in order to conduct inference. There are four assumptions that should be obtained which are 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. These assumptions can be checked from the residuals scatterplots which are generated as part of the multiple regression procedure. Residuals are the differences between the obtained and predicted dependent variable scores. If any of the statistical assumptions of the model are not met, then the model is not appropriate for the data. The normal distribution can be seen through histogram graph, plot P-P, plot Q-Q, kurtosis or skewness. Certain transformations of the data might be done when a model does not satisfy these assumptions, so that these

assumptions are reasonably satisfied for the transformed model (4).

2.4 Step 2: Selecting suitable methods of multiple linear regression

A great deal of care should be taken in selecting predictors for a model because the values of the regression coefficients depend upon the variables in the model. Therefore, the predictors included and the way in which they are entered into the model can have a great impact. In hierarchical regression predictors are selected based on past work and the experimenter decides in which order to enter predictors into the model. While in the standard regression or also called forced entry is a method in which all predictors are forced into the model simultaneously. In stepwise regression methods decisions about the order in which predictors are entered into the model are based on a set of statistical criteria. The three different versions of this stepwise regression method approach are forward selection, backward deletion and stepwise regression. SPSS allows us to decide for any one of these methods and it is important to select appropriate one (3).

2.5 Step 3: Interpreting the output

Through the output generated from SPSS, we can verify the values of Pearson coefficient, multiple coefficient of determination, R^2, multiple correlation coefficient, R and adjusted multiple coefficient of determination. A correlation is a measure of the linear relationship between variables. Value of strength for the correlation coefficient is as the following table [5].

Table 1 Value of Strength for Correlation Coefficient

Size of Correlation Coefficient (r)	Strength
0.91 to 1.00 or -0.91 to -1.00	Very strong
0.71 to 0.90 or -0.71 to -0.90	Strong
0.51 to 0.70 or -0.51 to -0.70	Medium
0.31 to 0.50 or -0.31 to -0.50	Weak
0.01 to 0.30 or -0.01 to -0.30	Very weak
0.00	No correlation

Multiple determination coefficient, R^2 is the proportion of the total variation in the *n* observed values of the dependent variable that is explained by the overall regression model [6]. By referring Minitab Methods and Formulas, the higher the R^2, the better the model fits of data. R is the positive squared root R^2 [7].

Adjusted R^2 is a measure of the loss of predictive power or shrinkage in regression. The adjusted R^2 will explain how much variance in the outcome would be accounted for if the model had been derived from the population from which sample was taken (3). By referring Minitab Methods and Formulas, the larger adjusted R^2, the better the model fits of data.

2.6 Step 4: Developing equation of multiple linear regression

In this research, the hypotheses are:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$$

$H_a :$ At least one of the $\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10}$ and β_{11} does not equal to 0 which says that,

$H_0 :$ 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 are not the response factor to the balance of trade.

$H_a :$ At least one of the 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 are the response factors to the balance of trade.

The model of multiple linear regression can be represented as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11} + e$$

where,

$\beta_0 =$ A constant, the value of y when all x values are zero

$\beta_i =$ The slope of the regression surface

(The β_i represents the regression coefficient associated with each x_i)

$y =$ Balance of trade

$x_1 =$ Exports of cocoa beans

$x_2 =$ Exports of crude petroleum

$x_3 =$ Exports of palm oil

$x_4 =$ Exports of plywood plain

$x_5 =$ Exports of sawn timber

$x_6 =$ Imports of fertilizers, manufactured

$x_7 =$ Imports of motor cars, completely built-up

$x_8 =$ Imports of petroleum products

$x_9 =$ Imports of refined beet and cane sugar

$x_{10} =$ Imports of rice

$x_{11} =$ Imports of tubes, pipes and fittings of iron or steel

$e =$ An error term, normally distributed about a mean of 0

(For purposes of computation, the e assumed to be 0)

3. RESULTS & DISCUSSION

Figure 1 shows the histogram and normal probability plot of the data. The histogram shows a roughly normal distribution (a bell-shaped curve). SPSS draws a curve on the histogram to show the shape of the distribution. The normal probability plot is nearly linear, most of the points fall along the straight line, indicating that the error terms are normally distributed. The straight line in the plot represents a normal distribution and the points represent the observed residuals.
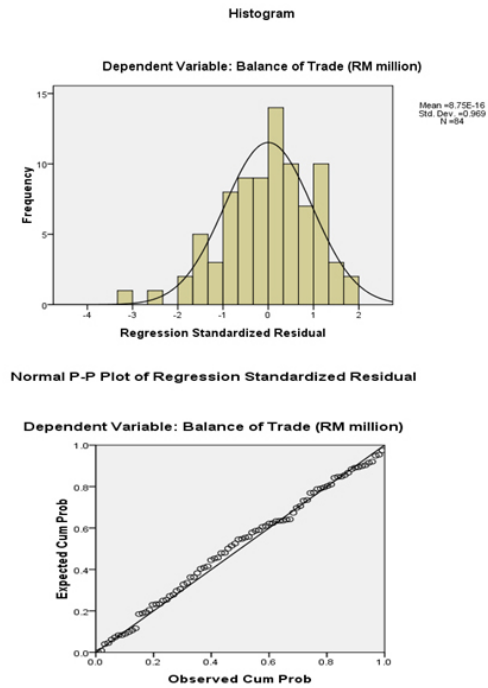

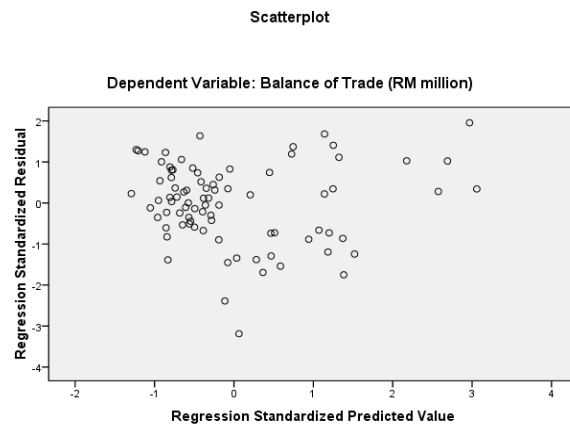**Fig. 1** Histogram and normal P-P plot of distributed residuals

Fig. 2 Plot of *ZRESID against *ZPRED

Figure 2 shows the plot of standardized residuals against standardized predicted values. The points are randomly and evenly dispersed throughout the plot. This

pattern is indicative of a situation in which the assumptions of linearity and homoscedasticity have been met.

Table 2 Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.812[a]	.659	.655	295.89105	
2	.865[b]	.748	.742	256.05568	
3	.916[c]	.839	.833	205.88418	
4	.922[d]	.850	.843	199.79567	
5	.928[e]	.861	.852	193.91434	1.143

a. Predictors: (Constant), Exports of Palm Oil (RM' 000)
b. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000)
c. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000)
d. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000), Imports of Motor Cars, Completely Built-up (RM' 000)
e. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000), Imports of Motor Cars, Completely Built-up (RM' 000), Exports of Plywood Plain (FOB RM' 000)
f. Dependent Variable: Balance of Trade (RM million)

Stepwise method has been chosen to use as the method of regression in this analysis. Durbin-Watson test has been performed in order to test the assumption of independence of errors. Durbin-Watson should be between 1.5 and 2.5 indicating the values are independent. From the Table 1, the value of Durbin-Watson is 1.143 where it indicates that the assumption of the error terms are independent has been met.

From the Model Summary in Table 2, the value of the multiple determination coefficient, R^2 is 0.659(Model 1) and it shows that there are 65.9% variation in the balance of trade is explained by export of palm oil. The value of R^2 is 0.748 (Model 2) and it shows that there are 74.8% variation in the balance of trade is explained by exports of palm oil and exports of crude petroleum. The value of R^2 is 0.839 (Model 3) and it shows that there are 83.9% variation in the balance of trade is explained by exports of palm oil, exports of crude petroleum and imports of petroleum products. The value of R^2 is 0.850 (Model 4) and it shows that there are 85.0% variation in the balance of trade is explained by exports of palm oil, exports of crude petroleum, imports of petroleum products and imports of motor cars. The value of R^2 is 0.861 (Model 5) and it shows that there are 86.1% variation in the balance of trade is explained by exports of palm oil, exports of crude petroleum, imports of petroleum products, imports of motor cars and exports of plywood plain. The regression equation appears to be very useful for making predictions since the value of R^2 for Model 5 is close to 1.

The adjusted R^2 gives us some idea of how well the model generalized and ideally we would like its value to be the same, or very close to, the value of multiple determination coefficient, R^2. The difference for the final model (Model 5) is small, 0.861 - 0.852 = 0.009 (0.9%). This shrinkage means that if the model were derived from the population rather than a sample it would account for approximately 0.9% less variance in the outcome.

By comparing all the models, Model 5 (R^2=0.861; adjusted R^2=0.852) is the best model fits to the data. This is because the higher the value of multiple correlation coefficient, R or multiple determination coefficient, R^2 and adjusted R^2, the better the model fits to the data.

Table 3 ANOVA Table

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.390E7	1	1.390E7	158.719	.000[a]
	Residual	7179224.143	82	87551.514		
	Total	2.108E7	83			
2	Regression	1.576E7	2	7882310.532	120.222	.000[b]
	Residual	5310725.259	81	65564.509		
	Total	2.108E7	83			
3	Regression	1.768E7	3	5894760.842	139.066	.000[c]
	Residual	3391063.797	80	42388.297		
	Total	2.108E7	83			
4	Regression	1.792E7	4	4480449.945	112.240	.000[d]
	Residual	3153546.543	79	39918.311		
	Total	2.108E7	83			
5	Regression	1.814E7	5	3628466.017	96.495	.000[e]
	Residual	2933016.239	78	37602.772		
	Total	2.108E7	83			

a. Predictors: (Constant), Exports of Palm Oil (RM' 000)
b. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000)
c. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000)
d. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000), Imports of Motor Cars, Completely Built-up (RM' 000)
e. Predictors: (Constant), Exports of Palm Oil (RM' 000), Exports of Crude Petroleum (FOB RM' 000), Imports of Petroleum Products (RM' 000), Imports of Motor Cars, Completely Built-up (RM' 000), Exports of Plywood Plain (FOB RM' 000)
f. Dependent Variable: Balance of Trade (RM million)

The result from the ANOVA table in Table 3 shows that there is a significant relationship between all the five predictor variables and the criterion variable at the 0.05 level of significance where $p < 0.05$.

There is a significant result for exports of palm oil [$F_{(1,82)}$=158.72, $p < 0.05$], while for the combination of exports of palm oil and exports of crude petroleum, the result is also significant [$F_{(2,81)}$=120.22, $p < 0.05$]. There is also a significant result for the combination of exports of palm oil, exports of crude petroleum and imports of petroleum products [$F_{(3,80)}$=139.06, $p < 0.05$]. The combination of exports of palm oil, exports of crude petroleum, imports of petroleum products and imports of motor cars shows a significant result too [$F_{(4,79)}$=112.24, $p < 0.05$]. As in the last model, it shows a significant result

for the combination of exports of palm oil, exports of crude petroleum, imports of petroleum products, imports of motor cars and exports of plywood plain [F(5,78)=96.50, $p < 0.05$]. Hence, the initial model significantly improved the ability to predict the outcome variable, but that the new model (with the extra predictors) was even better. It is because the *F*-ratio is more significant. Therefore, Model 5 has been chosen as the best regression model in order to build the best regression equation.

Table 4: Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-291.699	85.722		-3.403	.001
	Exports of Palm Oil (RM' 000)	.001	.000	.812	12.598	.000
2	(Constant)	-255.549	74.490		-3.431	.001
	Exports of Palm Oil (RM' 000)	.001	.000	.470	5.533	.000
	Exports of Crude Petroleum (FOB RM' 000)	.001	.000	.453	5.338	.000
3	(Constant)	-161.795	61.493		-2.631	.010
	Exports of Palm Oil (RM' 000)	.001	.000	.624	8.662	.000
	Exports of Crude Petroleum (FOB RM' 000)	.001	.000	.596	8.338	.000
	Imports of Petroleum Products (RM' 000)	-.001	.000	-.410	-6.730	.000
4	(Constant)	35.022	100.356		.349	.728
	Exports of Palm Oil (RM' 000)	.001	.000	.595	8.403	.000
	Exports of Crude Petroleum (FOB RM' 000)	.001	.000	.620	8.845	.000
	Imports of Petroleum Products (RM' 000)	.000	.000	-.373	-6.101	.000
	Imports of Motor Cars, Completely Built-up (RM' 000)	-.003	.001	-.113	-2.439	.017
5	(Constant)	-164.670	127.619		-1.290	.201
	Exports of Palm Oil (RM' 000)	.001	.000	.589	8.552	.000
	Exports of Crude Petroleum (FOB RM' 000)	.001	.000	.665	9.429	.000
	Imports of Petroleum Products (RM' 000)	-.001	.000	-.433	-6.734	.000
	Imports of Motor Cars, Completely Built-up (RM' 000)	-.003	.001	-.117	-2.608	.011
	Exports of Plywood Plain (FOB RM' 000)	.002	.001	.113	2.422	.018

a. Dependent Variable: Balance of Trade (RM million)

Through the output interpretation of ANOVA table in Table 3, Model 5 has been chosen as the best regression model in order to build the best regression equation. Hence, the β values in Model 5 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.

$$\hat{y} = -164.67 + 0.001x + 0.001x_2 - 0.001x_3 - 0.003x_4 + 0.002x_5$$

where,

y = Balance of trade

x = Exports of palm oil

x_2 = Exports of crude petroleum

x_3 = Imports of petroleum products

x_4 = Imports of motor cars, completely built-up

x_5 = Exports of plywood plain

For exports of palm oil (β = 0.001): This value indicates that as exports of palm oil increases by one unit, balance of trade increase by 0.001 units. Both variables were measured in thousands, therefore, for every RM1,000 more gained on exports of palm oil, an extra 0.001 thousand balance of trade (RM1) are generated. This interpretation is true only if the exports of crude petroleum, imports of petroleum products, imports of motor cars and exports of plywood plain are held constant.

For exports of crude petroleum (β = 0.001): This value indicates that as exports of crude petroleum increases by one unit, balance of trade increase by 0.001 units. Both variables were measured in thousands, therefore, for every RM1,000 more gained on exports of crude petroleum, an extra 0.001 thousand balance of trade (RM1) are generated. This interpretation is true only if the exports of palm oil, imports of petroleum products, imports of motor cars and exports of plywood plain are held constant.

For imports of petroleum products (β = -0.001): This value indicates that as imports of petroleum products increases by one unit, balance of trade decrease by 0.001 units. Both variables were measured in thousands, therefore, for every RM1,000 more spent on imports of petroleum products, an extra 0.001 thousand balance of trade (RM1) are deducted. This interpretation is true only if the exports of palm oil, exports of crude petroleum, imports of motor cars and exports of plywood plain are held constant.

For imports of motor cars (β = -0.003): This value indicates that as imports of motor cars increases by one unit, balance of trade decrease by 0.003 units. Both variables were measured in thousands, therefore, for every RM1,000 more spent on imports of motor cars, an extra 0.003 thousand balance of trade (RM3) are deducted. This interpretation is true only if the exports of palm oil, exports of crude petroleum, imports of petroleum products and exports of plywood plain are held constant.

For exports of plywood plain (β = 0.002): This value indicates that as exports of plywood plain increases by one unit, balance of trade increase by 0.002 units. Both variables were measured in thousands, therefore, for every RM1,000 more spent on exports of plywood plain, an extra 0.002 thousand balance of trade (RM2) are generated. This interpretation is true only if the exports of palm oil, exports of crude petroleum, imports of petroleum products and imports of motor cars are held constant.

4. CONCLUSION

In this study, a forecasting regression model has been developed through the multiple linear regression analysis. The best regression model to forecast the balance of the trade in the future is:

$$y = -164.67 + 0.001(exports\ of\ palm\ oil) + 0.001(exports\ of\ crude\ petroleum) - 0.001(imports\ of\ petroleum\ product) - 0.003(imports\ of\ motor\ cars) + 0.002(exports\ of\ plywood\ plain)$$

To conclude, the five significant predictors which are the exports of palm oil, exports of crude petroleum, imports of petroleum products, imports of motor cars, completely built-up and exports of plywood plain are equally important in predicting balance of trade. The assumptions seem to have been met and so it can probably be assumed that this model would generalize to any balance of trade being released.

REFERENCES

[1] Department of Industrial Development and Research, (2010, December 8), Retrieved from http://www.sabah.gov.my/didr/aLL_Links/Investment/ultimate/EconomicDev.htm.
[2] R. Peck, C. Olsen & J. Devore (2001), Introduction to statistics and data analysis, California: Duxbury Thomson Learning.
[3] A. Field (2003), Discovering statistics using SPSS for windows, London: SAGE Publications.
[4] J.E. Matson & B.R. Huguenard (2007), Evaluating aptness of a regression model, Journal of Statistics Education 15(2).
[5] Y.P. Chua (2009), Statistik penyelidikan lanjutan II: ujian regresi, analisis faktor dan analisis SEM, Kuala Lumpur: McGraw-Hill Education.
[6] B.L. Bowerman, R.T. O'Connell & A.B. Koehler (2005), Forecasting, time series, and regression, 4th edition, California: Thomson Brooks/Cole.
[7] R.I. Levin and D.S. Rubin (1994), Statistics for management, 6th edition, Englewood Cliffs, N.J.: Prentice Hall.