Stepwise Approach in Lagged Variables Time Series Modeling: A Simple Illustration

IOP Conf. Series: Materials Science and Engineering **621** (2019) 012009 doi:10.1088/1757-899X/621/1/012009

Stepwise Approach in Lagged Variables Time Series Modeling: A Simple Illustration

Yusma Yanti[1], **Septian Rahardiantoro**[2]

[1]Department of Computer Science, Pakuan University, Indonesia
[2]Department of Statistics, Bogor Agricultural University, Indonesia

Corresponding author: yusmayanti.fn@gmail.com

Abstract. Modeling approach in time series data commonly used to forecast the response variable based on the pattern of the predictor variables. The complicated cases occurred when in the model need the lagged variables from these variables. It can increase the number of predictor variables in the model. In this research, the increasing of number of predictor handled by the stepwise method in the regression analysis approach. All possible lag from the variables generated before the stepwise take action to choose the appropriate variables in the model. The illustration in this research based on the advertising-sales relationship of 36 consecutive months. Result of the model can determine the significance predictor variables in the model, and also can give the appropriate goodness of fit criteria for forecasting.

Keywords: forecasting, lagged variables, stepwise regression, time series regression.

1. Introduction

Time series modeling is a field of research that has the purpose of collecting and studying past observations over a period of time to develop appropriate models that describe the inherent structure of the series. The model obtained will be used to make a series forecast in the future [1]. Forecasting the time series can thus be called the act of predicting the future by understanding the past. Forecasting is very important to predict the future in some areas such as business, economics, finance, science and engineering, etc[2]. Because of this, in order to obtain the correct forecasting and in accordance with the expected model, it should have sufficient data for the time series underlying it. It is clear that successful time series forecasting depends on constructing the right model. In a lot of cases of time series modeling, the lagged variables of predictors or the response often needed to construct the model. In this situation can make the number of predictor variables in the model increase. Therefore, the model constructed become more complicated and difficult to understand. The method for reducing these predictor variables nacessary to take in this case one of the appropriate method is stepwise approach.

Stepwise regression is one of the methods to get the best model from a regression analysis. The first step is done by determining the correlation matrix between the dependent variable with the independent variable. The first variable entry is the variable with the highest correlation and significant with the dependent variable, the second incoming variable is the variable whose partial correlation is highest and still significant, after a certain variable goes into the model then the other variables in the model are evaluated, if any variable which is not significant then the variable is issued [1].

This research will apply the stepwise process to reduce the number of predictor variables from all the possible lagged predictor variables in the model. Data in this research reported in Blattberg and Jeuland (1981), that is advertising-sales relationship data cases. In the end of this research also obtain the

forecasting from the final model, and the evaluating how good the model constructed to forcast the sales variable.

2. Methods

The approach of this research would be determined in the process below: assume there are a response variable (y_t)and an independent variable (x_t), therefore the regression model that can construct [3]:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon \quad (1)$$

1. Generate all possible of lagged variable from response and predictor variable to be the predictor variables in the model. Let k number of lag to generate, therefore the predictor variables generated such as:

$$y_{t-1}, \dots, y_{t-k}, x_{t-1}, \dots, x_{t-k}$$

2. Construct the regression model using all lagged variables and predictor variable:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 x_{t-1} + \cdots + \beta_{k+1} x_{t-k} + \beta_{k+2} y_{t-1} + \cdots + \beta_{2k+1} y_{t-k} + \varepsilon \quad (2)$$

3. Apply the stepwise process using specified alpha to enter and alpha to remove.
4. Construct the regression model based on the predictor variables selected in the step 3 and evaluate the goodness of fit from these model.
5. Take the forecasting based on the model constructed in the step 4 and evaluate the forecasting goodness of fit using $MAPE$ and $RMSE$ with formula [4]:

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{e_t}{y_t}\right| \times 100\% \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2} \quad (4)$$

3. Data

The data in this research consist two variables with 36 consecutive months, which the study focus on the relationship between advertising and sales, reported in Blattberg and Jeuland (1981). The variables are 36 consecutive monthly sales (y_t) and advertising expenditures (x_t) of a dietery weight control product. Researchers in marketing have realized that the effect of advertising may not give effect in the same period in which the advertisement is seen.Therefore, the lagged variables needed in this model from the base model form:

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon$$

For the process of constructing the model, the 36 series of data seperated to be two parts. First part consists 30 series that called training data for constructing the best model. After the best model constructed, this model evaluated in the remaining 6 series of data that called testing data. The evaluation of forecasting by $MAPE$ and $RMSE$ using the testing data.

4. Result and Discussion

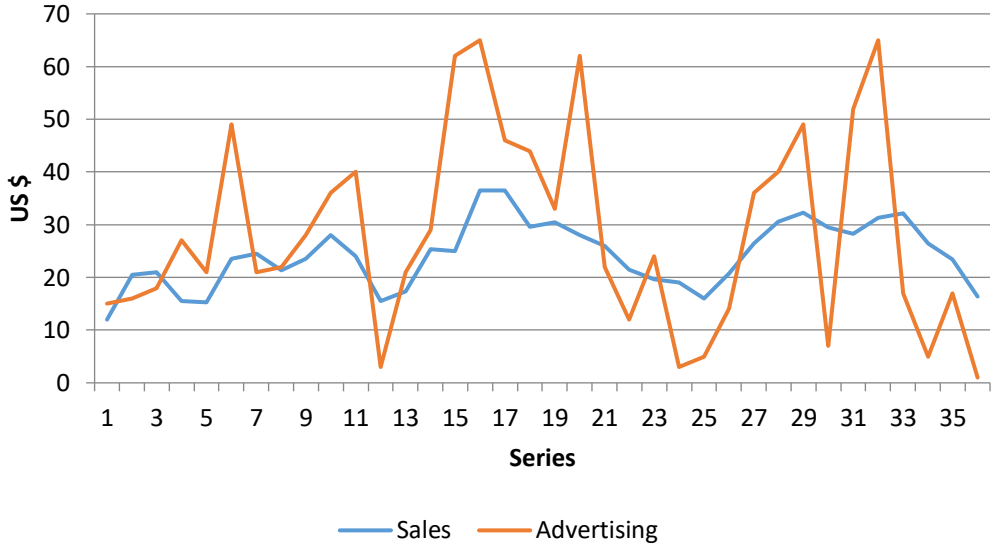

Figure 1. Time series plot of the data

Based on Figure 1 in general, it can be stated that with the increase in advertising costs, then sales will increase as well. The value of sales experienced irregular ups and downs from the first month to the 36th month. A very significant increase in sales occurred in the 12th month to the 17th month. But after that there was a long decline and then there was a spike in purchases again after the 25th month, although the value of sales did not reach the maximum value that occurred in the month-to- 16. When viewed from the graph of advertising costs, overall advertising costs used in a period almost doubled the cost of sales. High enough ads look greatly affect sales, this is evidenced by the maximum ad month-16 and the maximum sales at the same time.

For modeling purpose, lag of sales and advertising value need to be generated, in this case lag 1 until lag 6. Figure 2 and 3 describe the correlation value between sales and lag of sales, also lag of advertising. Correlation calculation results in Figure 2 obtained that the largest correlation between sales with sales lag 1 that is equal to 0.686. The value can interpret that between sales with sales lag 1 there is a considerable positive correlation. In other words, sales value of this time has strong relationship with the sales value in the one period past. As viewed overall, the correlation continues to decrease as the calculation increases with the next lag sale. In lag 6 sales there is a negative correlation value of -0.365.

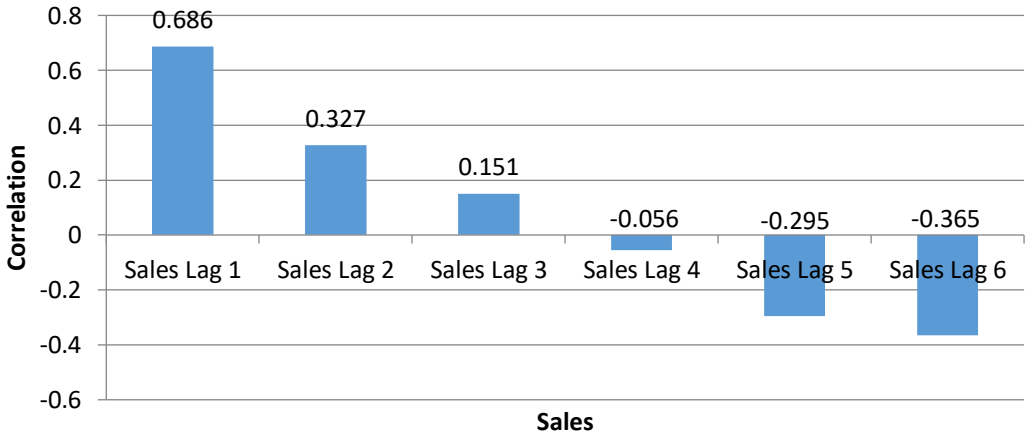

Figure 2. Correlation sales with lag of sales

Figure 3 demonstrates the correlation value between the sales and lag of advertising. It can be seen that the highest correlation occured between sales with advertising lag 1, 0674. The correlation value decreased along with the addition of advertising lag based on this figure. The big enough of positive correlation value also obtained between sales with advertisement that is 0.631.

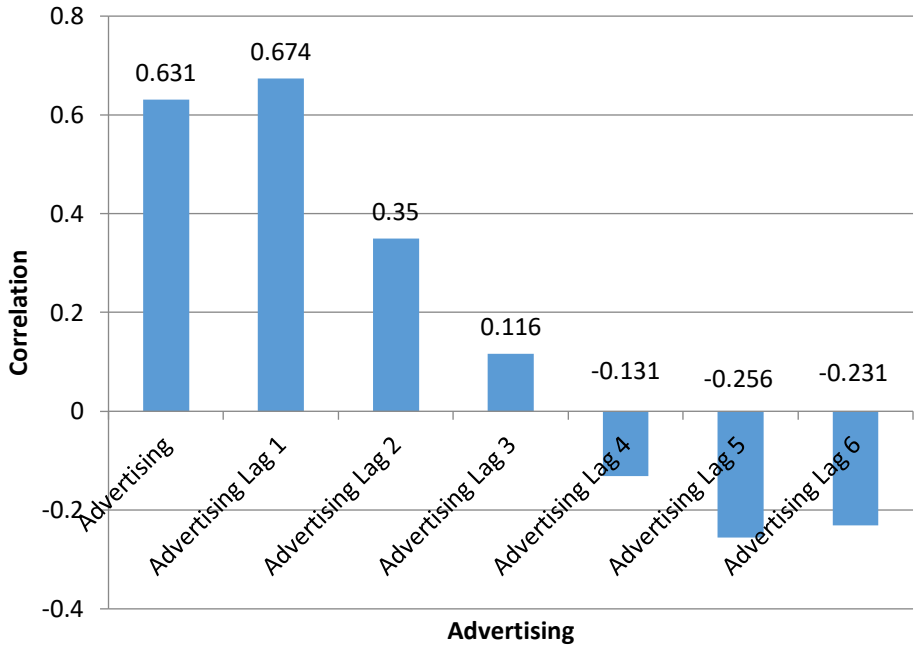

Figure 3. Correlation sales with lag of advertising

The next step, stepwise method applied in the data with the response variable of sales, and all possible predictor variables of lag of sales and advertising. The result of stepwise analysis can be seen in table 1:

Table 1. Stepwise Result

Step	Constant	Variables to enter				Goodness of fit			
		Advertising Lag 1	Advertising	Sales Lag 1	Advertising Lag 5	S	R-Sq	R-Sq(adj)	Mallows Cp
1	19.24	0.203 (0.000)				4.3	46.7	44.8	25.3
2	16.54	0.153 (0.000)	0.145 (0.000)			3.3	68.7	66.41	6.1
3	10.07	0.099 (0.014)	0.132 (0.000)	0.330 (0.020)		3.1	74.7	71.79	2.4
4	12.69	0.105 (0.007)	0.127 (0.000)	0.3 (0.024)	-0.064 (0.044)	2.9	78.6	75.12	0.7

Note: ( ) indicates p-value

In this analysis there are four steps. In the first step, variable enter in the model is advertising lag 1. Step two obtained advertising variables, step three obtained variable sales lag 1, and on step four obtained advertising lag 5. Based on the goodness of the model with the lowest S value of 2.87 and the value R-

Sq (adj) is the highest of 75.12, it can be concluded that this analysis is good enough. So the model chosen is model with predictor of advertising lag 1, advertising, sales lag 1, and advertising lag 5

The following is the result of regression modeling time series with stepwise method with predictor variable based on stepwise method.

Sales = 12.2 + 0.104 Advertising Lag 1 + 0.125 Advertising + 0.320 Sales Lag 1- 0.0621 Advertising Lag 5

Based on these results the four selected variables affect sales significantly. Consequently in this case the sales variable at time t is affected by advertising at time t-1, advertising, sales at time t-1 and advertising at time t-5 with R-Sq of 78.5. Using the model, the goodness of prediction using data testing obtained good results with RMSE of 4.555 and MAPE 13.149%.

In addition, the picture below in Figure 4 presented the forecasting pattern of the analysis. Based on this illustration, the result of regression modeling time series with predictor variable selection using stepwise method can be used as a good alternative to choose variable containing time series data.

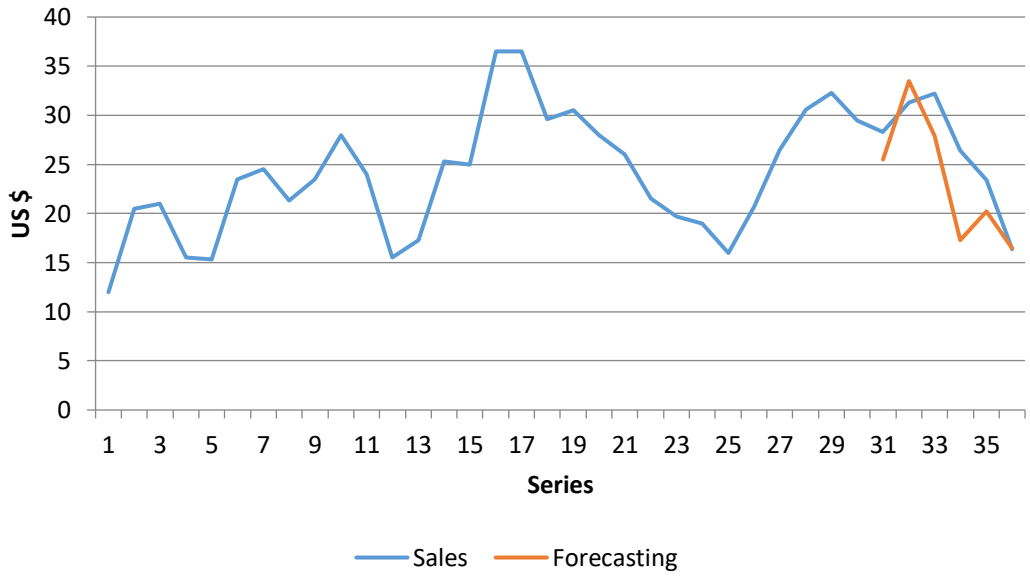

Figure 4. Time series plot of forecasting based on the final model

5. Conclusion

Based on the illustration presented in this study, the selection of predictor variables in the regression model using stepwise method on time series data gives a good result. Thus, although many predictor variables are generated based on the time lag, the researcher does not need to be binging to select which predictor variables are in the model. This approach seems good to reduce the number of predictor variables based on lagged variables generated.

References

[1] Wang, Mengchao, Wright J, Buswell R & Brownlee A 2013 A Comparison of Approaches to Stepwise Regression for Global Sensitvity Analysis Used with Evolutionary Optimalization. 13th Conference of International Building Performance Simulation Association, Chambéry, France, August 26-28

[2] Abraham, Bovas & Ledolter J 1981 *Statustical Methods for Forecasting*. John Wiley & Sons, Inc.

[3] Adhikari R & Agrawal K 2013 *An Introductory Study on Time Series Modeling and Forecasting*.

[4] Myers R & Milton J 1991 *A First Course in The Theory of Linear Statistical Models*. PWS-KENT Publishing Company