Journal of Automation and Control, 2015, Vol. 3, No. 3, 118-121
Available online at http://pubs.sciepub.com/automation/3/3/16

DOI:10.12691/automation-3-3-16

Regression Analysis and Seasonal Adjustment of Time Series

Eva Ostertagová[1,*], Oskar Ostertag[2]

[1]Department of Mathematics and Theoretical Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Nemcovej 32, 042 00 Košice, Slovak Republic
[2]Department of Applied Mechanics and Mechatronics, Faculty of Mechanical Engineering, Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic
*Corresponding author: eva.ostertagova@tuke.sk

Abstract The aim of this article is to demonstrate the dummy variables for estimation seasonal effects in a time series, to use them as inputs in a regression model for obtaining quality predictions. Model parameters were estimated using the least square method. After fitting, special tests to determine, if the model is satisfactory, were employed. The application data were analyzed using the MATLAB computer program that performs these calculations.

Keywords: *seasonal time series, dummy variables, trigonometric regression functions, method of least squares, residual analysis*

Cite This Article: Eva Ostertagová, and Oskar Ostertag, "Regression Analysis and Seasonal Adjustment of Time Series." *Journal of Automation and Control*, vol. 3, no. 3 (2015): 118-121. doi: 10.12691/automation-3-3-16.

1. Introduction

If we analyze the evolution of time series, we are interested not only in the main development trend of the indicators, but also in the course and intensity of any periodic fluctuations, which these time series present. When working with time series, the data must be adjusted seasonally. The aim of seasonal adjustment is to uncover the underlying dynamics in the development of the investigated phenomena and allow a direct comparison of their development in different seasons within the year. There are many methods of seasonal adjustment and their classification is not easy, because in practice the techniques used are a combination of several methods. Often they apply different types of moving averages, which eliminate from the time series the components the frequency of which does not exceed the number of observations forming the moving average length. To eliminate seasonal component regression methods based on the theory of linear regression model are also used. In case, where the nature of the seasonal component may change, e.g. the Winters exponential smoothing is applied.

2. Regression Approaches to the Seasonal Component of Time Series

In the construction of the forecasts of seasonal time series, a regression model with artificial (dummy) variables with simultaneously estimated trend and seasonality parameters can be used. Artificial variable is used to quantify the effect of the respective period on the estimated value of the investigated variables. The trend component is modeled via suitable regression function, for example line, parabola, and so on. The seasonal component is expressed using artificial (zero unit) variables that assign a value to the time series unit in case it is found in the considered season and zero otherwise.

Let us assume an additive time series model in which the value of the indicator y_t in the *t*-period is given by the sum $y_t = T_t + S_t + \varepsilon_t$, where T_t is the trend component, S_t is the seasonal component and ε_t is a random component. In the presence of free parameter (constant) in the model trend, in order to avoid multicollinearity, seasonality is modeled as a qualitative variable using the $s-1$ of artificial variables, where s is the length of the season included in the time series. Furthermore, we assume that the time series has a linear trend and quarterly seasonality. The relevant regression model can then be formulated for $t = 1, 2, \ldots, n$ in the form of

$$y_t = \beta_0 + \beta_1 t + \alpha_2 d_{t2} + \alpha_3 d_{t3} + \alpha_4 d_{t4} + \varepsilon_t, \qquad (1)$$

where the artificial variables are defined as vectors

$$\begin{aligned} d_{t2} &= (0, 1, 0, 0, 0, 1, 0, 0, \ldots), \\ d_{t3} &= (0, 0, 1, 0, 0, 0, 1, 0, \ldots), \\ d_{t4} &= (0, 0, 0, 1, 0, 0, 0, 1, \ldots), \end{aligned} \qquad (2)$$

of length of which is equal to n number of the time series of observations.

Since the artificial variable attains the value of one in a particular observation, we declare that in this period, to the value generated from a linear trend we shall add the value of seasonal fluctuations, which is calculated compared to the base period, which is in this case the first quarter of the year.

The artificial variable $d_{t1} = (0, 0, 0, 0, 0, 0, 0, 0, \ldots)$ is a zero vector and the effect of the first quarter is included in the intercept β_0 of the linear trend, which is interpreted in terms of the base level of the studied variables.

Model (1) contains a trend, seasonal and random component. Model parameters can be estimated using the least square method. The estimated model will take the form of:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\alpha}_2 d_{t2} + \hat{\alpha}_3 d_{t3} + \hat{\alpha}_4 d_{t4}. \quad (3)$$

The verification of the suitability of the regression model (1) is analogous to that in any other regression model. Particularly important is to test the heteroscedasticity and autocorrelation of the random component.

The estimated regression model (3) can be used for the construction of point and interval forecasts. Forecasting requires us to choose the time variables in the horizon of $h > 0$ and for the seasonal variables, substitute the unit values of the respective seasons in the horizon h.

In case of the regression model with artificial variables we shall adjust the estimated trend $\hat{T}_t$ and the seasonal factors I_j, $j = 1, 2, 3, 4$, to the form of [1]:

$$\hat{T}_t = (\hat{\beta}_0 + \bar{a}) + \hat{\beta}_1 t, \quad (4)$$

$$I_1 = -\bar{a}, I_2 = \hat{\alpha}_2 - \bar{a}, I_3 = \hat{\alpha}_3 - \bar{a}, I_4 = \hat{\alpha}_4 - \bar{a}, \quad (5)$$

where

$$\bar{a} = (\hat{\alpha}_2 + \hat{\alpha}_3 + \hat{\alpha}_4)/4 \quad (6)$$

is the "average" of seasonal regression parameters.

In an analogous manner we shall proceed in case of twelve month seasonality.

Another regression method for eliminating seasonal component is based on the fact that this component is estimated by means of a suitably selected mathematical function. The most commonly used are trigonometric functions with the period length equal to the number of periods s in the year, or a fraction of this number.

Provided that the trend of the considered time series is linear, the model may have e.g. this shape for $t = 1, 2, \ldots, n$:

$$y_t = \beta_0 + \beta_1 t + \beta_2 \cos(2\pi t/s) + \beta_3 \sin(2\pi t/s) + \varepsilon_t. \quad (7)$$

Since it is a general linear regression model, estimates of the parameters may be obtained by the least square method.

In case the coefficient of determination R^2 for the stated model is too small, we can continue to add to the model further unit values in the form of the considered trigonometric functions, with a half, fourth, or even smaller period, e.g. $\beta_4 \cos(4\pi t/s)$, $\beta_5 \sin(4\pi t/s)$, etc. From the models listed we select the one for which we achieved, for example, the maximum value of the coefficient of determination and also which best meets the other criteria imposed on the linear regression model.

3. The Application of Regression Models with artificial Variables and trigonometric Functions at Selected Time Series

We have data available on the number of sold pieces of selected articles of a business company engaged in the Internet sales of automotive accessories for individual quarters of the year, during the period of 2008 − 2014.

Figure 1 displays the time series presented in a form of plot via line chart.

The presented graph makes clear, that the stated time series has in the respective period an increasing, approximately linear trend and quarterly seasonality. The proposed regression model with artificial variables will have the form of (1).

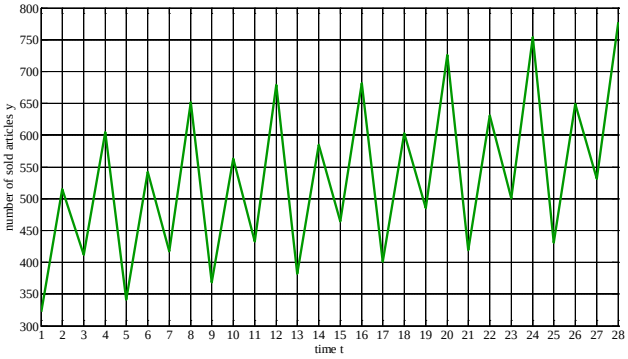

Figure 1. The development of the number of articles sold in the period of 2008-2014

The model estimated by the least squares method is:

$$\hat{y}_t = \hat{T}_t + \hat{S}_t = 308.5201 + 5.5424t + 197.8862 d_{t2} + 71.6295 d_{t3} + 299.3728 d_{t4}. \quad (8)$$

For two-sided 95% confidence intervals for regression coefficients applies, that:

$$\beta_0 \in \langle 298.7445, 318.2957 \rangle,$$

$$\beta_1 \in \langle 5.0682, 6.0166 \rangle,$$

$$\alpha_2 \in \langle 187.1461, 208.6262 \rangle,$$

$$\alpha_3 \in \langle 60.8580, 82.4009 \rangle,$$

$$\alpha_4 \in \langle 288.5493, 310.1963 \rangle.$$

To test the statistical significance of individual coefficients of the regression model the *t*-tests were used, where we received the following result values of test statistics and *p*-values:

[65.2874, 24.1790, 38.1151, 13.7565, 57.2181],

$[1.2863 \cdot 10^{-27}, 7.4491 \cdot 10^{-18}, 2.7269 \cdot 10^{-22}, 1.3820 \cdot 10^{-12}, 2.6265 \cdot 10^{-26}]$.

Since *p*-values are in all cases below the significance level $\alpha = 0.05$, all regression coefficients are considered statistically significant. The same result has also been provided by the confidence intervals for regression coefficients, since none of them contains zero value.

Based on the resulting value of the coefficient of determination $R^2 = 0.9953$ we can conclude that the model explained the variability of the number of units sold of selected articles to 99.53%.

The least squares method provides unbiased point estimates of parameters of the linear regression model while meeting certain assumptions about the probability distribution of random errors ε_t, for $t = 1, 2, \ldots, n$, within the model.

We assume, that the random errors ε_t [2,3]:

- have normal distribution,
- have zero mean values, i.e. $E(\varepsilon_t) = 0$;
- have constant variance (homoscedasticity), i.e. $V(\varepsilon_t) = \sigma^2$;
- are not correlated to each other (in case of the normality of the distribution are independent), i.e. the covariance $K(\varepsilon_i, \varepsilon_l) = 0$ for each $i \neq l$, $i, l = 1, 2, \ldots, n$.

The most important methods of regression model analysis include residual analysis. It is based on the assumption that the residuals e_t represent the point estimate of random errors ε_t. The equation $e_t = y_t - \hat{y}_t$ applies, i.e. (classical) residual is the difference of empirical and theoretical values.

The assumptions, on which the model is based, are generally verified by simple graphs, respectively, using known statistical tests [4,5].

In case of the graph displaying standardized residuals versus the theoretical values (see Figure 2), i.e. of a scatter plot $(\hat{y}_t, e_{St})$ applies, that the model is good if approximately 95% of residuals lies in the interval $(-2, 2)$. Also residuals have to be randomly distributed around zero and the plot must not show any indication of a potential trend or pattern of development [2,6].

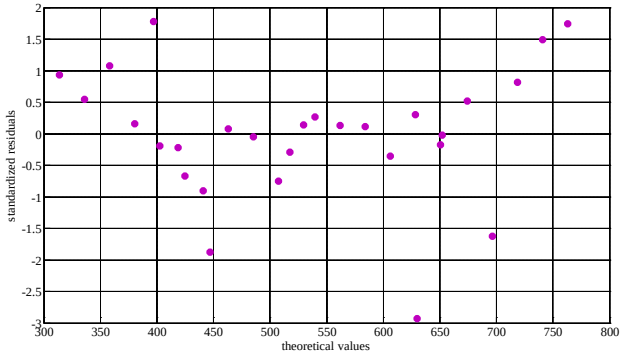

Figure 2. The dependence of standardized residuals on the theoretical values

The normality of the distribution of random errors we verified using the Anderson-Darling test of goodness-of-fit. On the significance level of $\alpha = 0.05$ we have tested the null hypothesis $H_0 : F(x) = F_0(x)$ against the alternative hypothesis $H_1 : F(x) \neq F_0(x)$, where $F(x)$ is the distribution function of random selection (residuals) and $F_0(x)$ is the distribution function of the normal distribution. We attained these results: Anderson-Darling statistics $AD = 0.5801$, p-value = 0.1217 > 0.05. Thus with a 95% reliability we can claim that random errors have a normal distribution.

Further, we tested the null hypothesis H_0: *random errors are uncorrelated* compared to the alternative hypothesis H_1: *random errors are correlated*. We have applied the Durbin-Watson test on the significance level of $\alpha = 0.05$ with the following results: statistics $DW = 2.6948$, whilst the p-value = 0.1068 > 0.05. We therefore do not reject the hypothesis on the no correlation of the random errors.

Based on these results it can be stated that model has good quality, and therefore it can be used to calculate extrapolations of the quarterly changes in the number of sold pieces of the selected articles of goods in the year 2015.

We shall use the equations (5) and (6). For the "average" of the seasonal regression parameters applies that $\bar{a} = 142.2221$.

For seasonal factors, we get the following results:

$$I_1 = -142.2221, I_2 = 55.6641, \\ I_3 = -70.5926, I_4 = 157.1507. \tag{9}$$

The presented results can be interpreted so as the average number of sold pieces of goods annually in the first quarter decreased by about 142 pieces, in the second quarter increased by about 56 pieces, in the third quarter decreased by about 71 pieces and in the fourth quarter increased by about 157 pieces compared to the trend.

Based on the equation (4) we get for the trend estimate the relation:

$$\hat{T}_t = 450.7422 + 5.5424t. \tag{10}$$

Figure 3 shows a plot of the stated time series and the estimated trend.

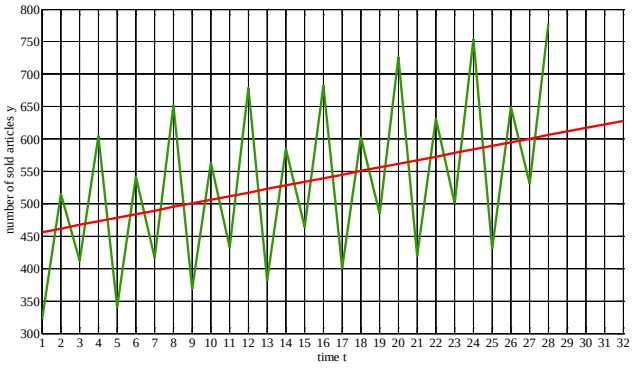

Figure 3. The plot of the stated time series together with the estimated trend

The extrapolated values of the original time series for each quarter of the year 2015 can be obtained based on the basis of the estimated model (8), respectively, on the basis of the respective seasonal factors (9) and the estimated trend (10).

The prediction for individual quarters of the year is as follows:

$$\hat{y}_{29} = 308.5201 + 5.5424 \cdot 29 = 469.2500,$$

or $\hat{y}_{29} = 450.7422 + 5.5424 \cdot 29 - 142.2221 = 469.25,$

$$\hat{y}_{30} = 308.5201 + 5.5424 \cdot 30 + 197.8862 = 672.6786,$$

or $\hat{y}_{30} = \hat{T}_{30} + I_2 = 672.6786,$

$$\hat{y}_{31} = 308.5201 + 5.5424 \cdot 31 + 71.6295 = 551.9643,$$

or $\hat{y}_{31} = \hat{T}_{31} + I_3,$

$$\hat{y}_{32} = 308.5201 + 5.5424 \cdot 32 + 299.3728 = 785.25,$$

or $\hat{y}_{32} = \hat{T}_{32} + I_4.$

Now follow the application of the linear regression model (7). The good model estimated by the least squares method is:

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 t + \hat{\beta}_2 \cos\frac{\pi t}{2} + \hat{\beta}_3 \sin\frac{\pi t}{2} + \hat{\beta}_4 \cos \pi t, \quad (11)$$

where $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_4$ are unbiased estimators of the true regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$.

Least squares parameter estimates for this model are $\hat{\beta} = (450.7422, 5.5424, 50.7433, -35.8147, 106.4074)^T$.

The predictions for individual quarters of the year 2015 are the same as in the case of application of the regression model with artificial variables: $\hat{y}_{29} = 469.25$, $\hat{y}_{30} = 672.6786$, $\hat{y}_{31} = 551.9643$, $\hat{y}_{32} = 785.25$. The coefficient of determination is in this case $R^2 = 0.9953$ too.

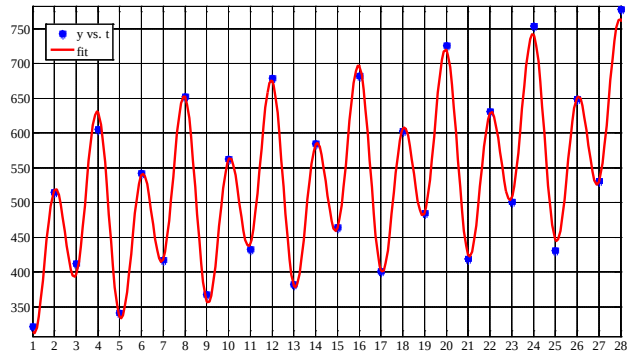

Figure 4. The plot of the measured data with the estimated trend (11)

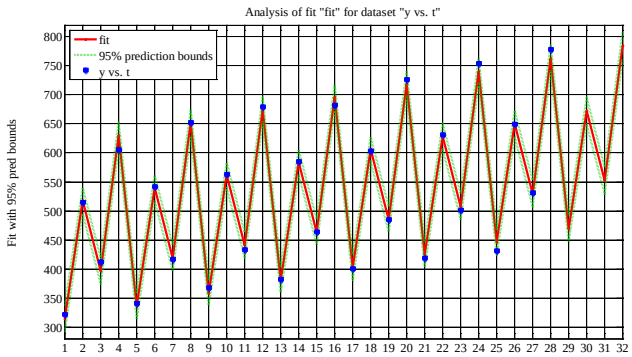

Figure 5. The plot of the stated time series together with the estimated trend (11) and 95 % prediction interval

Figure 4 presents a scatter diagram of the measured data with the least squares fitted trend (11). Figure 5 shows more so the 95 % prediction interval for sold pieces of selected articles of a business company.

4. Conclusion

The current paper presents the analysis of time series with linear growing trend and additive seasonal component. To determine the seasonal component, a method based on the theory of linear regression model with artificial variables, i.e., variables that are discrete or qualitative in nature, so they cannot be directly quantified, was used. For eliminating seasonal component was used regression model with trigonometric functions too.

The analysis of the seasonal component allows us to increase our knowledge about the patterns of behavior of a given effect, respectively phenomenon, and contribute to the construction of better forecasts of the considered time series.

Acknowledgement

This work was supported by the VEGA grant scheme no. 1/1205/12 Numerical Modeling of Mechatronic Systems and the VEGA grant scheme no. 1/0393/14 Analysis of Causes of Mechanical System Failures by the Quantification of Strains and Stress Fields.

References

[1] Arlt, J., Arltová, M., Rublíková, E., *The Analysis of Economic Time Series with Examples* (in Czech), VŠE Prague, 2002.

[2] Montgomery, D.C., Runger, G.C., *Applied Statistics and Probability for Engineers*, John Wiley & Sons, 2003.

[3] Ostertagová, E., *Modelling Using Polynomial Regression*, Procedia Engineering, 48 (2012), p. 500-506.

[4] Ostertagová, E., *Applied Statistics* (in Slovak), Elfa, Košice, 2011, 161 pp.

[5] Ostertagová, E., *Applied Statistics in the Computational Environment of the MATLAB software* (in Slovak), TU Košice, 2015, 175 pp.

[6] Ostertagová, E., Ostertag, O., *Time Series Modelling*, The 4th International Conference on Modelling of Mechanical and Mechatronic Systems, Technical University of Košice, Slovak Republic, Proceedings of conference,2011, p. 380-384.