THE DETERMINATION OF A COMPLEX SUMMARY OF STATISTICAL METHODS FOR THE BEST FITTED REGRESSION OR TREND FUNCTION SELECTION

Vladimíra Hovorková Valentová – Kateřina Gurinová

Abstract

One of the problems which can arise during the process of looking for the best fitted regression or trend model with the help of various kinds of statistical methods is the fact that the conclusions resulting from the different approaches application are not in correspondence in many cases. This problem comes forth in the courses of statistics, for example. Basic courses of statistics present the best fitted function selection with the help of overall F test, t tests, coefficient of determination and some other statistics. In master degree students can learn other methods, especially Durbin-Watson test. But the results of Durbin-Watson test are often in disharmony with tradicionally used F test and t tests and students are then confused and they are not able to make a clear conclusion of the process of the best fitted function selection. The aim of this paper is a creation of the complex procedure for the best fitted regression or trend function selection which allow us to combine various aspects and procedures leading to the best fitted model determination. The paper also includes a manual for the situations when the results of various kinds of statistical methods do not comply.

Key words: Durbin-Watson test, F test, regression function, t tests, trend function

JEL Code: C18, C22

Introduction

Examining the suitability of regression and trend functions is a complex process that involves different statistical methods. The main aim is to describe the relationships, dependencies, and correlations among the variables as closely as possible and define them mathematically. The selected regression and trend function, in an ideal situation, should closely reflect the dependencies and relationships among the examined variables. The strength of such dependency is not only connected with the quality assessment of the estimated model but also with the assessment of the implied relationships actual existence.

Since the regression and trend parameters values are not observable, there is no other possibility but to make estimation based on the sample data. Yet, it should be noted that the estimation error of the dependent variable value for any independent variable value with the help of the selected model is caused by a number of reasons. For example, it could be an inappropriate regression parameter estimation method, an insufficient representative sample, an inadequate data collection, or the influence of other factors on the dependent variable. The decision that the best function of an unknown regression or trend function approximation is the linear function could be problematic as the dependence, in reality, is not linear (Hušek, 1999).

1 Selection of Regression or Trend Functions Best Fitted Model

There is no doubt that the choice of the best fitted model should, in the first place, result from an objective analysis of the issues; in other words, the choice should be supported by the theoretical knowledge of the analysed variables relationships. However, if a suitable model cannot be selected in a priori, it is necessary to consider the adhesion of the individual functions to the selected data through various tests and characteristics.

1.1 Estimation of Linear Model Parameters

According to the most widely used classical approach, the model parameters are considered unknown constants, which are to be estimated by a suitable method. The most frequently used method is the least square method based on the minimalization of the residual sum of squares. The first step is to specify a linear model, which complies with not only the classical requirements for the least square method but also with the condition of data normality. The merit of the individual independent variables can be verified with the help of t test and the overall significance of the regression model with the help of F test (Seger, Hronová, Hindls, 1998).

An important prerequisite for the linear regression model estimation with the help of the classical LSM is, among others, zero covariances. This prerequisite is not often fulfilled, especially when estimating the model parameters from time series. If a random model component in selected time series is correlated with a random component or components from previous time series, then we talk about autocorrelation or serial correlation of random components. Autocorrelation is not understood as dependence among variables but as dependence among value sequence of one variable, which are arranged in time.

1.2 Assessment of the Selected Model Quality Based on a Residuals Analysis

Residuals, which are calculated based on the selected model, are a suitable diagnostic tool for assessing the selected model quality. Their non-random arrangement signals drawbacks of the applied model. Due to the fact that residuals do not automatically estimate good values of an irregular component, it is not possible to make meaningful statistical judgements about regression or trend function parameters without probability assumptions of a random component. The increasing sample size diminishes the negative impact of the random component assumption violation, yet it is not possible to ignore probability assumptions even in big sample sizes (Hebák et al., 2013).

Suitable methods help to verify whether the obtained residuals are really an estimation of a random component or not. To do so, they have to be mutually independent variables. Therefore, apart from the zero mean assumption, it is necessary to define the variation of irregular components assumption and the components mutual dependency.

The most frequently used assumption is the hypothesis of homoscedasticity of random errors, which assumes that random errors with zero mean values have a constant variance in time, and are mutually independent (Black, 2010). In case the above-mentioned requirements are fulfilled, the sequence of random errors creates so-called white noise. Another assumption of a random variable is the hypothesis of heteroscedasticity of random errors, where random errors with zero mean values are considered mutually independent with variable variance.

The assumption of random error autoregression is a frequent assumption of a random component. It is based on the assumption that each random error consists of a component dependent on the previous error and a random component (Ostrom, 1978). The coefficient of the autocorrelation of the neighbouring random errors is constant and the random component is a sequence of mutually independent random errors with zero mean values and constant variances (Bowerman, O Connell, 1997). The causes of autocorrelation are numerous, e.g. an incorrect or inaccurate specification of a model’s mathematical form, an inclusion of the dependent variable measurement errors into a random component of the model, or the usage of lagged independent variables.

1.2 Durbin-Watson Test

A wide range of tests, which are based on the properties of the stated residuals, can be used to verify the assumptions of a random component. Durbin-Watson test is a very frequently used method which verifies whether random errors are really independent (Durbin, Watson, 1950).

The null hypothesis formulates the assumption of random error independence, the alternative hypothesis, on the other hand, assumes the dependence, which is expressed by an autoregressive scheme. The test statistic DW value ranges from zero to four. An approximate evaluation of the test statistics is frequently used in practice. The test statistic values around the number two point at the random errors independence; values close to zero point at the positive autocorrelation, and values close to the number four point at the negative autocorrelation. Special tables, which include critical test values for the given number of observations,
a number of model parameters and various significance levels, are used to conduct an accurate evaluation. The test results are in a certain area equivocal and this area decreases with the increasing sample size (Tillman, 1975).

Durbin-Watson test cannot be used to test autocorrelation of higher orders, not even in the case of a nonlinear form of random components autocorrelation. Moreover, the test power function weakens when the model includes lagged dependent variables or a stochastic variable as independent variables. Durbin suggested, for these cases, a modified, asymptotic, first order autocorrelation test, based on the statistics h (Durbin, 1969).

The first order autocorrelation DW test can also be used while testing incorrect model specifications. A low value of Durbin-Watson test does not necessary mean a positive first order autocorrelation, but can indicate a mistake in the model specifications. For example, an omission of an important independent variable is reflected in the increase of residuals and their positive correlation. An indication of this is a fact that a sequence of several positive residuals regularly alternates with a sequence of several negative residuals (Durbin, Watson, 1951).

The consequences of random components autocorrelation are similar to heteroscedasticity. The estimates of linear regression model parameters via the classical least square method remain impartial and consistent, but they do not have a minimal variance and are not even asymptotically efficient. The estimates of variance and standard deviations are biased, therefore, it is neither possible to rely on the calculated confidence intervals nor on the usual tests which weaken their power function. The positive autocorrelation usually leads to the fact that the estimates of estimated parameters standard errors are underestimated, therefore, they incline to zero.

2 Choosing a Best Fitted Model – Didactic Approach

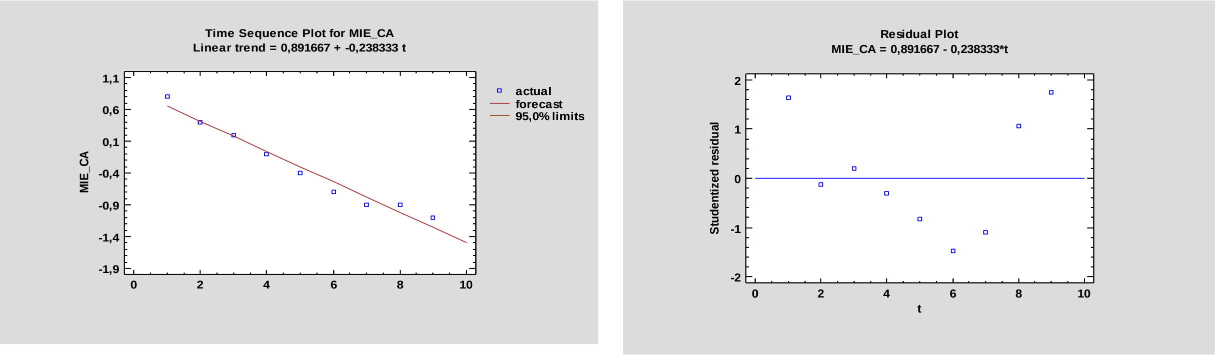
As apparent from the above-mentioned facts, the process of looking for a best-fitted type of regression and trend functions is a very complex matter, which causes great complications while teaching basic statistic courses. The basic courses do not provide students with the detailed knowledge of the whole complex issue. The adopted procedure is that the basic courses students are introduced to the issue of looking for a best fitted type of regressive and trend function with mainly the help of F test, t test, and R-squared, or the mean squared error.

Only in following courses of statistics in the master studies is the students' knowledge widened by further information and methods such as Durbin-Watson test. However, it sometimes happens that the results of the previously lectured methods do not correspond with the information the newly taught methods bring. For example, the conclusions of Durbin-Watson test conflict with the traditionally used F test and t test, which leads to the fact that students are unable to provide an unambiguous conclusion of the whole process.

2.1 Real Data Study

Let us show here which kind of cases student can meet in the process of his/her research. We work with data about three year average of current account balance as % of GDP in 2004 – 2012 (in following text just CA). This data are obtained from MIP Scoreboards which are the part of the Alert Mechanism Report which is regularly created by the Commission to the European Parliament, the Council, the European Central Bank and the European Economic and Social Committee. It seems that it is possible to approximate this time series by linear trend – see Figure 1. This linear trend is given by the equation: $\hat{T}_t = 0{,}892 - 0{,}238t$.

Fig. 1: Time Sequence Plot for the CA indicator (Linear Trend) and Residual Plot



Source: own (with the help of STATGRAPHICS Centurion XVII)

The choice of the trend function is then reexamined by overall F test which is supplemented by t tests on specific regression coefficients. Let us suppose the 5% significance level for all the tests. The results of the tests are in Tables 1 and 2.

Tab. 1: Outcomes of t-tests (Linear Trend)

Parameter	Estimate	P-Value
Constant	0.891667	0.0000
Slope	-0.238333	0.0000

Source: own calculation in the programme STATGRAPHICS Centurion XVII

Tab. 2: Outcomes of ANOVA (Linear Trend)

Source	Sum of Squares	Df	F-Ratio	P-Value
Model	3.40817	1	213.33	0.0000
Residual	0.111833	7		

Source: own calculation in the programme STATGRAPHICS Centurion XVII

Outcomes presented in Tables 1 and 2 should make a student to make a conclusion that the used linear trend function is the best model. But when we look at the results of Durbin-Watson test, it is clear that it is not possible to make a definite conclusion. DW statistic is equal to 0.883 and P-Value equals to 0.0062. Thus, Durbin-Watson test leads to the null hypothesis rejection which signifies that the random errors are not independent. The question is what can be the reason of such a result and how to solve it. Firstly, we can verify the main assumptions of the model use. One of the basic assumptions of linear model is a normality of data distribution. We tested it with the help of Shapiro-Wilk test. The value of W statistic is 0.939 and P-Value equals to 0.5616. So, we can suppose that the data are normally distributed. The assumption of errors homoscedasticity is also valid – the value of the F statistic is 0.103 and P-Value equals to 0.0938. Often it is possible to see the reason of the negative result of Durbin-Watson test in the residual plot – see Figure 1. In this specific example the autocorrelation is probably caused by the middle part of the residuals series.

It is also necessary to check if another trend function will be better for the data approximation. If we look at the list of simple functions (with two parameters) which the programme STATGRAPHICS Centurion XVII contains, no one is suitable. The result of Durbin-Watson test is also for all of them negative, it means that it leads to the null hypothesis rejection. If we use for this time series approximation the parabola, we find that all needed assumptions are held. The value of the DW statistic is 2.023 and P-Value 0.1589. The results of t tests and F test can be found in Tables 3 and 4. The quadratic trend equation is: $\hat{T}_t = 1.195 - 0.404x + 0.017x^2$.

As Tables 3 and 4 show all the tests are significant. The general movement of the CA indicator was not so simple to be possible to describe it with the help of some simple function with two parameters.

Tab. 3: Outcomes of t-tests (Quadratic Trend)

Parameter	Estimate	P-Value
Constant	1.19524	0.0000
t	-0.403918	0.0001
t^2	0.0165584	0.0051

Source: own calculation in the programme STATGRAPHICS Centurion XVII

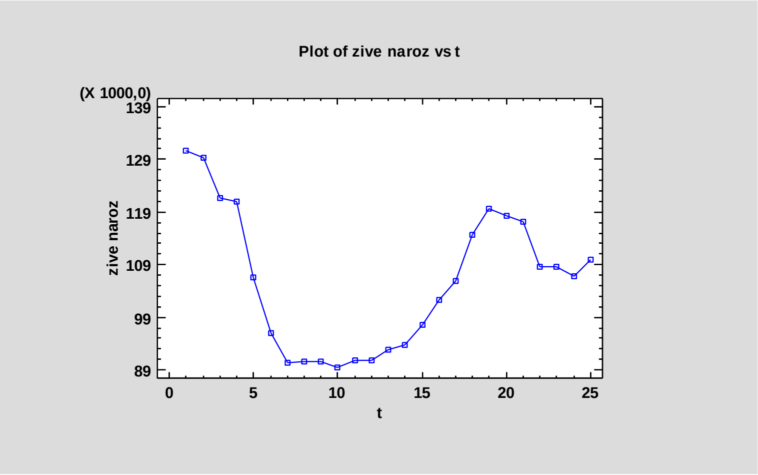
Tab. 4: Outcomes of ANOVA (Quadratic Trend)

Source	Sum of Squares	Df	F-Ratio	P-Value
Model	3.49261	2	382.61	0.0000
Residual	0.0273853	6		

Source: own calculation in the programme STATGRAPHICS Centurion XVII

More complicated situation arises when the use of the model with more than two parameters will not bring a satisfactory solution. Let us have the data about number of live births in the Czech Republic in 1990 – 2014 (see Figure 2). When we look at the Figure 2, it is clear that the use of one of the simple trend functions will not be suitable. If we approximate the data with the help of cubic polynomial, the outcomes of t tests and F test look very good – see Tables 5 and 6.

Fig. 2: X-Y Scatterplot – Live Births in the Czech Republic in 1990 – 2014



Source: own (with the help of STATGRAPHICS Centurion XVII)

Tab. 5: Outcomes of t-tests (Cubic Polynomial Trend)

Parameter	Estimate	P-Value
Constant	155659.0	0.0000
t	-16090.4	0.0000
t^2	1222.97	0.0000
t^3	-26.2743	0.0000

Source: own calculation in the programme STATGRAPHICS Centurion XVII

Tab. 4: Outcomes of ANOVA (Cubic Polynomial Trend)

Source	Sum of Squares	Df	F-Ratio	P-Value
Model	$3.59313*10^9$	3	45.66	0.0000
Residual	$5.50892*10^8$	21		

Source: own calculation in the programme STATGRAPHICS Centurion XVII

But when we carry out the Durbin-Watson test, we will find that the value of the DW statistic is 1.152 and P-Value 0.0002. Thus, Durbin-Watson test does not bring a positive result. In this case it is possible to suppose the existence of the cycle component in the time series and it can be the reason why the mathematical functions fail.

Discussion

The process of looking for the best regression and trend functions with the help of various statistical methods is considerably complicated because the application of various procedures often leads to ambiguous results. While some methods prefer one type of function, others can lead to another model selection. It depends on a certain data file very much and also on the choice of model selection criterions. At the same time the selection of model assessment procedures is dependent on the individual priorities of the problem solver. Thus, it is clear that it is not possible to determine the unambiguous procedure a priori. Due to these reasons it is quite hard to decide which methods should be included in basic statistical courses. Time for teaching the methods leading to a best regression/trend model selection is very limited and it is impossible to teach students all needed procedures and all connections between the methods. Therefore, it is necessary to improve students´ knowledge in other statistical courses which are taught especially in the master degree of study.

Conclusion

As resulting from presented theoretical facts and also from the real data study it is not possible to create the general procedure usable for any situation. The features of data are the important factor which is necessary to take into account before application of any statistical method. It is possible to apply the correct methods leading to a successful model selection just if we make a complete assessment of the real base of analysed process. The sample size is also belongs to considerable factors which have impact on the regression or trend function selection. When we suppose work with regression models, it is important the correct sampling method use.

When we work with real data files, we can distinguish a few situations which can appear. One of them is apparent from the first example mentioned above. The graph in Figure 2 as well as F test and t tests show the suitability of linear trend function use but Durbin-Watson test does not confirm this conclusion. But we can solve this situation using another trend function which Durbin-Watson test corresponds with results of F test and t tests.

We can meet more complicated situation when it is not possible to find any model which would have positive results of F test and t tests as well as Durbin-Watson test. If we suppose the time series analysis, the problem can be solved using some of the adaptive methods instead of classical approach. As mentioned above the assessment of real features of analysed data is very important because it can alert to a possibility of seasonal or cyclical variation occurrence. Just this fluctuating can cause that Durbin-Watson test indicates the errors dependency. Therefore, it is necessary to search if the periodical variations replacement leads to better results.

When we suppose work with regression models the procedures mentioned above are not usable in general. It is possible to recommend complete exploration of all the relevant assumptions above all, it means the data normality and errors homoscedasticity. The breaking of one of the assumptions can explain the conflict of tests. If we do not find any assumption violation, e. g. the incorrect method of data sampling can be the reason of the tests disagreement. The truth is that this problem cannot be solved in the phase of data analysis and we are not often able to verify this fact. Also insufficient sample size can be another reason why the tests give us conflicting results. It can have a negative impact on the conclusions influencing from a best model criterions use.

Acknowledgment

This paper was written in the frame of work on the MŠMT project, OP Education for Competitiveness, section of the support 2.3, called Innovation of the Study Programme Economics and Management with the Focus on Knowledge Economics, No. CZ.1.07/2.200/28.0317.

References

Black, K. (2010). *Business statistics: For contemporary decision making*. Hoboken, NJ: Wiley.

Bowerman, B. L., & O'Connell, R. T. (1997). *Applied statistics: Improving business processes*. Chicago: Irwin.

Durbin, J. (1969). Tests for serial correlation in regression analysis based on the periodogram of least-squares residuals. *Biometrika*, *56*(1), 1-15.

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrika*, *37*(3/4), 409.

Durbin, J., & Watson, G. S. (1951). Testing for Serial Correlation in Least Squares Regression. II. *Biometrika*, *38*(1/2), 159.

Hebák, P. (2013). *Statistické myšlení a nástroje analýzy dat*. Praha: INFORMATORIUM.

Hušek, R. (1999). *Ekonometrická analýza: Předmět a metody: Simulační modely a techniky: Ekonometrické prognózování*. Praha: Ekopress.

Chen, Y. (2016). Spatial Autocorrelation Approaches to Testing Residuals from Least Squares Regression. *PLOS ONE PLoS ONE*, *11*(1).

Ostrom, C. W. (1978). *Time series analysis: Regression techniques*. Beverly Hills, CA: Sage Publications.

Seger, J., Hronová, S., & Hindls, R. (1998). *Statistika v hospodářství*. Praha: ETC.

Tillman, J. A. (1975). The Power of the Durbin-Watson Test. *Econometrica*, *43*(5/6), 959.

Wei, W. W. (1990). *Time series analysis: Univariate and multivariate methods*. Redwood City, CA: Addison-Wesley Pub.

European Commission - Economy and Financial Affairs - MIP. (n.d.). Retrieved April 21, 2016, from http://ec.europa.eu/economy_finance/indicators/economic_reforms/eip/sba/index.cfm

Oficiální stránky Českého statistického úřadu. (n.d.). Retrieved April 26, 2016, from https://www.czso.cz/csu/czso/4-obyvatelstvo-

Contact

Vladimíra Hovorková Valentová
Technical University of Liberec
Studentská 1402/2, 461 17 Liberec 1
vladimira.valentova@tul.cz

Kateřina Gurinová

Technical University of Liberec

Studentská 1402/2, 461 17 Liberec 1

katerina.gurinova@tul.cz