INFO-1058 UDK: 519.6:51:007
Primljeno/Received: 2007-03-13 Izvorni znanstveni rad / Original Scientific Paper

REGRESSION ANALYSIS AND APPROXIMATION BY MEANS OF CHEBYSHEV POLYNOMIAL

REGRESIVNA ANALIZA I APROKSIMACIJA POMOĆU ČEBIŠEVLJEVIH POLINOMA

Nikola Tomašević, Marko Tomašević, Tatjana Stanivuk*
Faculty of Economics, University of Belgrade, Belgrade, Serbia*; Faculty of Maritime Studies, University of Split, Split, Croatia
Ekonomski fakultet, Sveučilište u Beogradu, Beograd, Srbija*; Pomorski fakultet, Sveučilište u Splitu, Split, Hrvatska

A b s t r a c t

The paper deals with the regression model, describes the procedure of getting regression coefficients and gives the analysis of the model variance. Chebyshev polynomials and their properties are also stated. The transformation of the regression model, from segment [a, b] to segment [-1, 1] is performed, as well as the approximation of the obtained regression polynomial, using prespecified accuracy polynomials of lower degree.

S a ž e t a k

Rad se bavi regresivnim modelom, opisuje postupak dobivanja regresivnih koeficijenata, te daje analitički model varijance. Promatraju se i Čebiševljevi polinomi i njihova svojstva. Izvodi se transformacija regresivnog modela od segmenta [a, b] do segmenta [-1, 1], kao i aproksimacija izvedene regresije polinoma, koristeći unaprijed određene polinome nižeg stupnja.

Introduction

Trend, as a notion, implies the expected development (or a development tendency, as it is frequently called) of a phenomenon in a future period, based on the observation of that phenomenon over a previous and long enough period. More precisely, on the basis of phenomenon development data in the previous period, i.e. a series of data in previous consecutive periods, a mathematical equation, best corresponding to the data given, is selected. The equation is then "extended" to future periods, assuming that the phenomenon will be developing in the same way, and its future values are calculated. The values usually refer to a sufficiently close future period.

Most frequently, the selected functions for expressing regression are the following:

- The linear regression in the form of:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \,, \quad i = 1, 2, ..., n$$

- Regression second degree polynomial (parabolic regression) in the form of:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^{\,2} \,, \quad i = 1, 2, ..., n$$

- Regression k degree polynomial ( ), $\beta_0, \beta_1, ..., \beta_k$ in the form of:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^{\,2} + ... + \hat{\beta}_k x_i^{\,k} \,, \quad i = 1, 2, ..., n$$

The calculation of function parameter values, which are used to express regression, is performed on the basis of the data already known from the previous period. The technique of determination of the coefficients itself is performed using the method of least squares. The parameters of the required function are selected to make the sum of square difference of known real values and the values of the required function minimum. We determine the value of the required coefficients by minimum, from these equations.

Regression analysis

Mutual connections among changeable phenomena can be divided into two groups: deterministic and stochastic. A deterministic connection, which is also called functional, refers to the case in which only one value of the independent

variable X corresponds to only one value of the dependent variable Y. This type of connection can be shown in a general expression: $Y = g(X)$, where g (X) is a function of X. A stochastic connection refers to the case in which whole series of possible values of the dependent variable correspond to only one value of the independent variable. The essence of the stochastic connections to be observed in this paper is the functional connection which exists between individual values of the independent variable X and average values of the dependant variable Y (i.e. expected values).

Thus:

$$E[Y] = f(X)$$

The main problem of regression is to estimate the functional $(x_1, Y_1),...,(x_n, Y_n)$ dependence on the basis of sample, as well as to the give statistic evaluation of accuracy of such estimations.

The regression $Y = Y(x)$ model consists of two additive parts: deterministic and stochastic. Hence, the regression model can be observed as the following sum:

$$Y = g(X) + e_i i = 1, 2, ..., n$$

The specification of the regression model as a stochastic model does not only imply its mathematical expression. It also implies assumptions which ensure the optimal estimation of parameters. The following assumptions are most frequently introduced:

1. Linearity – There is a linear connection between individual values of the independent variable, X, x_i and the corresponding Y value, and also the average value $E[Y_i]$.
2. $E(e_i) = 0$
3. Homogeneity of variance – Stochastic members have the equal deviation, or more precisely, equal variances. Hence:

$$V[e_1] = V[e_2] = ... = V[e_n] = \sigma^2.$$

4. There is no autocorrelation between e_i, $i = 1, 2, ..., n$, hence, there is no linear connection between any of two stochastic members e_i and e_j.
5. X is not a random variable (therefrom, the independent variable in the model is denoted with the small letter). This assumption indicates that the values of the independent variable are fixed. In other words, they must be chosen in advance by the researcher before sample taking.
6. e_i has a normal distribution so it can be written as: , i.e. the stochastic member has a normal distribution with the arithmetic mean $e_i \sim \rightarrow(0, \sigma^2)$ equivalent to zero and σ^2 variance.

Our aim is to find the best possible coefficient estimations of the regression model on the basis of the sample. The method of finding coefficient estimations of the regression model on the basis of the sample, $(x_1, Y_1),...,(x_n, Y_n)$ which will be applied here, is the method of least squares.

The basic regression model is the k degree polynomial in the form of:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + ... + \beta_k x_i^k + e_i, i = 1, 2, ..., n$$

where:

Y_i – the i dependant variable (project value)
x_i – i value of the independent variable
$\beta_0, \beta_1, ..., \beta_k$ – regression parameters (regression coefficient)
e_i – stochastic member or random error
n – sample size

The system of equations in its matrix notation is

$$Y = X\beta + e$$

where the following has been set:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 & x_1^2 & ... & x_1^k \\ 1 & x_2 & x_2^2 & ... & x_2^k \\ ... & ... & ... & ... & ... \\ 1 & x_n & x_n^2 & ... & x_n^k \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ ... \\ e_n \end{bmatrix}$$

In this notation, Y stands for the vector of the empirical values of the dependent variable, X stands for the matrix of the value of the regression variable X, β stands for the vector of the unknown parameter values, and e stands for the vector of the unknown values of the random variable e_i.

$$E[e_i] = 0, \quad \textit{per each } i$$

On the basis of the sample, the best possible estimations $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ with the corresponding regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ have to be found in order to determine the regression curve of the sample:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta}_k x_i^k, i = 1, 2, ..., n$$

In this sample the sign $\hat{Y}_i$ denotes Y value which is positioned on the best adjusted regression curve and therefore is called the adjusted value of Y.

As a rule, the regression curve of the basic set differs from that of the sample because the estimated values $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ differ from the real parameter values. The vertical deviation $\beta_0, \beta_1, \ldots, \beta_k$ (the difference) between the real value y_i and the adjusted value $\hat{y}_i$ is called the residual value and it is denoted by e_i:

$$e_i = y_i \quad \hat{y}_i = y_i \quad \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta}_k x_i^k\right),$$

The main idea of the method of least squares is to choose the regression curve with the least sum of squares residue out of all regression curves possible.

Thus, the minimum of the expression

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} y_i - \left[\left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta}_k x_i^k\right)\right]^2 \text{ has to be found.}$$

In the above expression $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are unknown. The procedure for minimizing is performed by finding the partial derivatives of the expression above by ??? and their equalization to zero. In this way, the following system of $k + 1$ equations is obtained. These equations are called normal equations:

$$\begin{gathered}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^{k+1} = \sum_{i=1}^{n} x_i y_i \\
\text{---------------------} \\
\hat{\beta}_0 \sum_{i=1}^{n} x_i^k + \hat{\beta}_1 \sum_{i=1}^{n} x_i^{k+1} + ... + \hat{\beta}_k \sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} x_i^k y_i
\end{gathered}$$

Their matrix notation is the following:

$$\left(X^T X\right) \cdot \hat{\beta} = X^T Y, \qquad \text{i.e.} \qquad \hat{\beta} = \left(X^T X\right)^{-1} \cdot \left(X^T Y\right),$$

where

$$X^T X = \begin{bmatrix} n & \sum x_i & ... & \sum x_i^k \\ \sum x_i & \sum x_i^2 & ... & \sum x_i^{k+1} \\ ... & ... & ... & ... \\ \sum x_i^k & \sum x_i^{k+1} & ... & \sum x_i^{2k} \end{bmatrix}; \qquad X^T Y = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ ... \\ \sum x_i^k y_i \end{bmatrix}.$$

If the assumptions of the regression model are fulfilled, the estimations obtained by the method of least squares will be the best, i.e. the most efficient and impartial linear estimations. Thus, the following is valid:

$$E\left[\hat{\beta}_0\right] = \beta_0, \quad E\left[\hat{\beta}_1\right] = \beta_1, ..., E\left[\hat{\beta}_k\right] = \beta_k,$$

i.e. the estimations $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_k$ are equal, on average, to the unknown parameters with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The standard deviation is denoted by $\sigma_{\hat{\beta}_i}$ and it is called the standard error of the estimation $\hat{\beta}_i$. It measures the estimation of $\hat{\beta}_i$ deviation from parameters β_i (i=1, ... , k) .

In order to consider the components of variability of the dependant variable, the sample regression drawn in the diagram (as in Figure 1) has to be observed. The diagram shows an arbitrary empirical value y_i, taken from the sample which corresponds to the value of the dependant variable x_i .

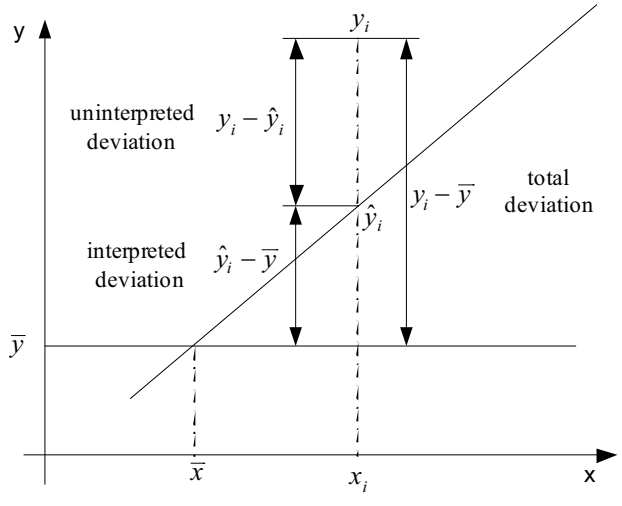

Figure 1

The total deviation of the dependant variable can be treated as the sum of interpreted and uninterpreted deviation :

$$\left(y_i - \bar{y}\right) = \left(\hat{y}_i - \bar{y}\right) + \left(y_i - \hat{y}_i\right).$$

$y_i - \bar{y}$ - total deviation
$\hat{y}_i - \bar{y}$ - interpreted deviation
$y_i - \hat{y}_i$ - uninterpreted deviation

This equality will be valid if we square and sum both sides by all values of the sample, so it can be said that the total variability equals to the sum of interpreted and uninterpreted variability:

$$e_i = y_i - \hat{y}_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 x_i^2 + ... + \hat{\beta}_k x_i^k\right)$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

i.e. $ST = SP + SR$

ST - total sum of squares
SP - interpreted sum of squares
SR - uninterpreted sum of squares

It is known that the sample variance is obtained when the sum of squares is divided by the number of degrees of freedom $\left(n-k-1\right)$ The estimated variance of regression model is obtained by the following formula:

$$s^2 = \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-\left(k+1\right)},$$

or in matrix notation:

$$\hat{\sigma}^2 = \frac{Y^T Y - \hat{\beta}^T\left(X^T Y\right)}{n-\left(k+1\right)}$$

The standard regression error is the square root of the residual variance and denotes the standard deviation of random error:

$$s = \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-\left(k+1\right)}} = \sqrt{\frac{Y^T Y - \hat{\beta}^T\left(X^T Y\right)}{n-\left(k+1\right)}}.$$

The standard error of the regression model parameters is given by the formula:

$$se\left(\hat{\beta}_j\right) = \hat{\sigma} - \sqrt{s_{jj}}, j = 1, 2, ..., k$$

where s_{jj} stands for diagonal matrix $\left(X^T X\right)^{-1}$ elements.

The following text will present the procedure for obtaining individual elements in the equation decomposition of sum of squares of deviation of empirical values of the dependant variable from the average value of the same variable.

Starting from the following formula:

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(y_i^2 - 2\bar{y}y_i + \bar{y}^2\right) = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2,$$

that is:

$$ST = Y^T Y - n\bar{Y}^2;$$

$$\sum_{i\ 1}^{n}\left(\hat{y}_i-\overline{y}\right)^2=\hat{\beta}_0\sum_{i\ 1}^{n}y_i+\hat{\beta}_1\sum_{i\ 1}^{n}x_iy_i+\ldots+\hat{\beta}_k\sum_{i\ 1}^{n}x_i^{\,k}y_i-n\overline{y}^2,$$

i.e.

$$SP=\hat{\beta}^T\left(X^TY\right)-n\overline{Y}^2,$$

hence:

$$SR=Y^TY-\hat{\beta}^T\left(X^TY\right)$$

The respective number of degrees of freedom with the total sum of squares is , with the interpreted part k and the residual sum of squares $n-(k+1)$.

The method of variance analysis examines whether the regression parameter of, for example, k- degree reduces significantly the residual sum of squares. In other words, it examines if the k -degree polynomial would be optimal.

Approximation using Chebyshev polinomials

In the analysis of the variance, when finding the optimal degree of the polynomial and testing its coefficient, it is very important for the polynomial to be of the least degree possible, because the mathematical calculation and search for theoretical data using the optimal polynomial may result in errors. These errors might accumulate and turn into system errors, which may lead to wrong conclusions. This has been the reason for choosing Chebyshev polynomials to reduce the polynomial degree and accelerate the process of search for theoretical data, keeping at the same time the coefficients given within the estimated intervals. In finding solutions to this problem, some areas of special functions have been used and they will be stated in the following pages.

Namely, it is not difficult to prove that, using the linear transformation

$$x=\frac{b-a}{2}t+\frac{b+a}{2},$$

polynomials

$$\tilde{T}_m\left(x\right)=\left(\frac{b-a}{2}\right)^m T_m\left(t\right)=\left(\frac{b-a}{2}\right)^m T_m\left(\frac{x-\frac{b+a}{2}}{\frac{b-a}{2}}\right)$$

keep at the segment $[a,b]$ all properties of Chebyshev polynomials $T_m(t)$ defined at the segment $[-1,1]$. It is also possible to prove that:

$$\max_{a\le x\le b}\left(\tilde{T}_m\left(x\right)-0\right(=\tilde{E}_m=\left(\frac{b-a}{2}\right)^m\cdot\frac{1}{2^{m-1}}=2\left(\frac{b-a}{4}\right)^m$$

Since

$$T_n\left(x\right)=\cos\left(n\arccos x\right),$$

and, if starting from the identity

$$\cos\left(n+1\right)a+\cos\left(n-1\right)a=2\cos na-\cos a$$

we put $a=\arccos x$, we get :

$$T_{n+1}\left(x\right)=2xT_n\left(x\right)-T_{n-1}\left(x\right)\left(-1\le x\le 1\right)$$

This relation can be applied to Chebyshev polynomial formation. Indeed, starting from

$T_0\left(x\right)=1$ and $T_1\left(x\right)=x$, and using (1) we get:

$$T_2\left(x\right)=2x\cdot T_1-T_0=2x^2-1,$$

$$T_3\left(x\right)=2x\cdot T_2-T_1=4x^3-3x,$$

$$T_4\left(x\right)=2x\cdot T_3-T_2=8x^4-8x^2+1$$

This procedure can be further extended. Using the above formulas, we shall approximate the polynomial:

$$f\left(x\right)=a_0+a_1x+a_2x^2+a_3x^3+a_4x^4$$

by the third degree polynomial, so that:

$$\max_{-1\le x\le 1}\left|f\left(x\right)-P_3\left(x\right)\right|<0,05.$$

Using the expressions above we have:

$$1 = T_0,\ x = T_1,\ x^2 = \frac{1}{2}\left(T_0 + T_2\right)$$

$$x^3 = \frac{1}{4}\left(3T_1 + T_3\right),\ x^4 = \frac{1}{8}\left(3T_0 + 4T_2 + T_4\right),$$

hence:

$$f\left(x\right) = a_0T_0 + a_1T_1 + \frac{a_2}{2}\left(T_0 + T_2\right) + \frac{a_3}{4}\left(3T_1 + T_3\right) + \frac{a_4}{8}\left(3T_0 + 4T_2 + T_4\right) =$$

$$= \left(a_0 + \frac{a_2}{2} + \frac{3a_4}{8}\right)T_0 + \left(a_1 + \frac{3a_3}{4}\right)T_1 + \left(\frac{a_2 + a_4}{2}\right)T_2 + \frac{a_3}{4}T_3 + \frac{a_4}{8}T_4.$$

Since $\left|T_n\left(x\right)\right| = \left|\cos\left(n\arccos x\right)\right| < 1$, and with the last member estimated under the condition (2), then we have to obtain:

$$\left|\frac{a_4}{8}\right| < 0,05$$

Consequently, for all fourth degree polynomials to which , the following is valid:

$$f\left(x\right) = \left(a_0 + \frac{a_2}{2} + \frac{3a_4}{8}\right)T_0 + \left(a_1 + \frac{3a_3}{4}\right)T_1 + \left(\frac{a_2 + a_4}{2}\right)T_2 + \frac{a_3}{4}T_3 =$$

$$= \left(a_0 + \frac{a_2}{2} + \frac{3a_4}{8}\right) + \left(a_1 + \frac{3a_3}{4}\right)x + \left(\frac{a_2 + a_4}{2}\right)\left(2x^2 - 1\right) + \frac{a_3}{4}\left(4x^3 - 3x\right) =$$

$$= \left(a_0 - \frac{a_4}{8}\right) + a_1x + \left(a_2 + a_4\right)x^2 + a_3x^3$$

Example: A capital investment intended for the production of series A based on the 15 quarterly values of production (variable x in comb.) is being analysed. Variable y (in 000 kn) denotes total investments. Here we have:

x_i- quarterly value of production in kunas for i = 1,2,…,15

y_i- total investment of dedicated production of series A, expressed in 000 kn.

The regression fourth degree polynomial is chosen as the total investments model. Using the method of least squares, parameters based on empirical values are estimated, and estimation standard errors are calculated. The regression equation is the following:

$$\hat{y} = 2434 + 85,7x - 0,03x^2 + 0,00004x^3 - 0,000002x^4$$

$$R^2 = 0,999\ .$$

Since $|a_4| = 0,000002 < 0,4$, the assumption is fulfilled, hence the obtained regression polynomial can be approximated by a third degree polynomial. Consequently, we have:

$$P_3(x) = 2434 + 85,7x - (0,03 + 0,000002)x^2 + 0,00004x^3 =$$

$$= 2434 + 85,7x - 0,030002x^2 + 0,00004x^3$$

We have obtained the third degree polynomial which approximates the regression fourth degree polynomial with prespecified level of significance and its coefficients still remaining within the calculated confidence intervals.

Conclusion

This paper presents the regression model analysis and variance analysis. When finding the optimal degree of a polynomial, and testing its coefficients in the variance analysis, it is very important to find a polynomial of the least degree possible as the regression polynomial of a higher degree may result in a higher order matrix, which would make matrix multiplication and the search for the inverse matrix more complicated. The results of special functions, i.e. Chebyshev polynomials, have been used in the paper. These polynomials approximate the polynomial obtained by the regression analysis to a lower degree polynomial with required accuracy, keeping the polynomial coefficient within the estimated intervals at the same time. The advantages of the result achieved are obvious because the method enables an easier and faster regression coefficient calculation. In addition, it avoids unnecessary difficulties in methodology of searching regression values.

Literature

1. Boyd, J. P.: *Chebyshev and Fourier Spectral Methods*, University of Michigan, Dover Publications, 2000.
2. Demeyer, J.: *Recursively enumerable sets of polynomials over a finite field*, J. Algebra 310, 2007.
3. Ditlevsen, O., and Madsen, M. O., *Structural reliability methods*, Technical University of Dermark, Lyngby, (2nd ed.), 2003.
4. Eisentr, K.: *Hilbert's tenth problem for algebraic function fields of characteristic 2*, Pacific J. Math. 210, 2003.
5. Sarapa, N., *Teorija vjerojatnosti* , Školska knjiga, Zagreb, 2002.
6. Tomašević, M., *Matematički i vektorski račun*, Pomorski fakultet Sveučilišta u Splitu, 2007.
7. Tomašević, M., Ristov, P., and Stanivuk, T., *Metodologija znanstvenoistraživačkog rada, Statističke metode u istraživanju*, Pomorski fakultet Sveučilišta u Splitu, 2007.
8. Vasilyev, N. and Zelevinsky, A. *A Chebyshev Polyplayground: Recurrence Relations Applied to a Famous Set of Formulas*. Quantum 10, 20-26, Sept/Oct.1999.
9. Vujanović, N., *Teorija pouzdanosti tehničkih sistema*, Vojnoizdavački novinarski centar, Beograd 1990.
10. http://mathworld.wolfram.com/ChebyshevPolynomialoftheFirstKind.html (accessed 2009-07-08)