# SEGMENTED REGRESSION BASED ON THEPRINCIPLE OF CUT-OFF POLYNOMIALS WITH A SOLVED ECONOMIC EXAMPLE

**MILOŠ KAŇKA**
University of Economics, Faculty of Informatics and Statistics
Department of Mathematics, Ekonomická 957, Prague 4, 148 00, Czech Republik
email: kanka@vse.cz

**Abstract**

*This article deals with segmented regression that is based on the so called cut-off polynomials of the first, or second, respectively third order. Most of the attention is given to the inference of the system of the normal equations gained by minimisation of the criterion of the least squares method. Certain attention is also given to the calculation of the unknowns of such a system with one of the Gauss methods. Parametric equations of the output regressive curve in the given model and the determination indexes of the observed variables are inferred. Those indexes also enable to assess the "quality" of each model. The author of this article also wrote a program for the three types of segmented regression. This program helps solve the aforementioned phases in much faster and simpler way.*

***Key words:*** *Cut-off polynomials, system of normal equations, indexes of determination of observed values.*

***JEL Codes:*** *C10, C63, C65*

***DOI: 10.15611/amse.2017.20.19***

## 1. Introduction

The time sequence of a comparable quantity does not generally develop smoothly in the given time period, but into several stages with varying dynamics instead. With the help of segmented regression we can quantify these stages better.
Segmented regression can be mathematically based on a so called cut-off polynomials, which are of a type of so called polynomial splines. For example, Makarov, V. L., and Chlobystov, V. V. (1983, 4-35 pp.) talk about this topic in their book.
Seger, J. (1988, 431-436 pp.) briefly talks about the importance of segmented regression and its use in economic matter, however it does not deal with its mathematical basics. Meloun, M., and Militky, J. (1994, 752-759 pp.) are talking about the following in their book - the base of segmented regression is solved in greater detail, from both points of view of B-spline basis functions, and cut-off polynomials (those have more pages dedicated to it than the B-spline in this book). This paper continues on this matter. The important formulas for the calculation of the elements of matrix of normal equations, and the elements of column vectors on the right sides, are deducted in an easier way. The system of those equations, after the calculation of given unknowns, gives us the parametrical equations of output regressional curve, which is the main solution of the calculation of the given problem.
The method I used seems to me to be the most appropriate in this case. The results, that are obtained, are given in chapter 2 - Applications of regression models, in paragraphs A, B, and C.

Now, let me follow up briefly about the segmented regression based on a principle of cut-off polynomials.

In space $R^m$ ($m > 1$, integer) with canonical basis $e_s = \delta_{sj}$ (Kronecker's delta; $s = 1, 2, \ldots, m$, $j = 1, 2, \ldots, m$), we generally consider different points $P_i = x_j^{(i)}$ ($i = 1, 2, .., n$, $j = 1, 2, \ldots, m$; index $j$ is here, and in the following text represents numbers 1, 2, …, $m$). Elsewhere, we consider the so called nodal points on real axis $t$:

$T_1 < T_2 < \ldots < T_k$, where integer $k \geq 1$ and with the others, so called complementary nodal points $T_0 < T_1$ a $T_{k+1} > T_k$; we will call these points the main nodal points $T_1 < T_2 < \ldots < T_k$.

In interval $<T_{l-1}, T_l>$, $l = 1, 2, \ldots, k+1$, in which the $t$ is changing, we consider the increasing sequence of points $T_{l-1} \leq t_{l,1} < t_{l,2} < \ldots < t_{l,n(l)} < T_l$, where integer $n(l) \geq 1$ indicates the number of those points. We assume, at the same time, that every point $t_{l,w}$ ( = $t_{lw}$, if the record is understandable), exactly one of the above mentioned points $P_i$ is assigned. Then the following applies $n = \sum_{l=1}^{k+1} n(l)$. In union $\cup_{l=1}^{k+1} <T_{l-1}, T_l>$ we will consider a real function of variable $t$ of shape

$$g_j(t) = \gamma_j^{(1)} + \gamma_j^{(2)} t + \ldots + \gamma_j^{(Q+1)} t^Q + \sum_{r=1}^{k} \gamma_j^{(r+Q+1)} [(t - T_r)_+]^Q , \qquad (1)$$

where $\gamma_j^{(1)}, \gamma_j^{(2)}, \ldots, \gamma_j^{(k+Q+1)}$ are real parameters, which means linear (for $Q = 1$), quadratic (for $Q = 2$) and cubic (for $Q = 3$) cut-off polynomial. By the symbol $(x)_+$ we mean a real function of a variable $x$:

$$(x)_+ = \begin{cases} x & for \quad x>0 , \\ 0 & for \quad x \leq 0 . \end{cases}$$

We will assume that the process which is conditioned by the points $P_i$ and to those points associated values of variable $t$ is additive, which means that for all $j$, $l$, $w$, which come into consideration, the following is true:

$$x_j^{(lw)} = g_j(t_{lw}) + \varepsilon_j^{(lw)} , \qquad (2)$$

where $\varepsilon_j^{(lw)}$ are independent and identically distributed random variables with the constant variance. The estimates $c_j^{(1)}, c_j^{(2)}, \ldots, c_j^{(k+Q+1)}$ of parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, \ldots, \gamma_j^{(k+Q+1)}$ are possible to obtain based on the minimisation of the criterion of the method of least squares (with the use of (1), (2))

$U_j = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} [x_j^{(lw)} - g_j(t_{lw})]^2 = \sum_{l=1}^{k+1} \sum_{w=1}^{n(l)} [ - x_j^{(lw)} + \sum_{q=1}^{Q+1} \gamma_j^{(q)} t_{lw}^{q-1} + +$
$\sum_{r=1}^{k} \gamma_j^{(r+Q+1)} [(t_{lw} - T_r)_+]^Q]^2 = U_j (\gamma_j^{(1)}, \gamma_j^{(2)}, \ldots, \gamma_j^{(k+Q+1)})$;

since this moment we will denote this $[(t - T_r)_+]^Q$ in a simpler form $(t - T_r)_+^Q$.

The next progress leading to the so called system of normal equations (in a matrix format)

$$\underline{\underline{M}} \cdot \underline{c}_j = \underline{\underline{Z}}_j \qquad (3)$$

for the calculation of vector of estimations $\underline{c}_j = (c_j^{(1)}, c_j^{(2)}, \ldots, c_j^{(k+Q+1)})^T$ of vector of parameters $(\gamma_j^{(1)}, \gamma_j^{(2)}, ..., \gamma_j^{(k+Q+1)})^T$ is mentioned in an article by Kaňka, M. (2016, pp. 60-72). Alongside formulas for the calculation of elements in the matrix $\underline{\underline{M}} = (m_{pq})_{1 \leq p, q \leq k+Q+1}$ and column vectors, are $\underline{\underline{Z}}_j = (z_{pj})_{1 \leq p \leq k+Q+1}$.

By solving the system of equations (3), we obtain estimates $c_j^{(1)}, c_j^{(2)}, ..., c_j^{(k+Q+1)}$ of parameters $\gamma_j^{(1)}, \gamma_j^{(2)}, ..., \gamma_j^{(k+Q+1)}$ of cut-off polynomial (1) in interval $<T_0, T_{k+1}>$, and for $j = 1, 2, \ldots, m$ are equations

$$x_j = G_j(t) = c_j^{(1)} + c_j^{(2)} t + \ldots + c_j^{(Q+1)} t^Q + \sum_{r=1}^{k} c_j^{(r+Q+1)} (t - T_r)_+^Q , \; t \in <T_0, T_{k+1}> \qquad (4)$$

parametrical equations of so called regressive curve in space $R^m$, which is an output of regression model of the considered tracking process. For $Q = 1$ we call this curve *linear* regressive curve, for $Q = 2$ *quadratic* regressive curve, and for $Q = 3$ *cubic* regressive curve. We are usually satisfied with those three types.

## 2. Applications of regression models

We will base on the values mentioned in the table 1. The task is the execution of a regression survey for $Q \in \{1, 2, 3\}$. We shall proceed within the meaning of the text in theoretical part of this paper.

Table 1: Time series analysed in the regression models

| Year, quarter | | t | Export of goods in milliards of CZK $x_1$ | Import of goods in milliards of CZK $x_2$ | Unemployment rate in percent $x_3$ |
|---|---|---|---|---|---|
| 2010 | $Q_4$ | 0 | 683.902 | 660.274 | 6.9 |
| 2011 | $Q_1$ | 3 | 703.550 | 648.436 | 7.2 |
| | $Q_2$ | 6 | 722.784 | 680.140 | 6.7 |
| | $Q_3$ | 9 | 696.451 | 660.761 | 6.5 |
| | $Q_4$ | 12 | 755.905 | 698.227 | 6.4 |
| 2012 | $Q_1$ | 15 | 787.325 | 692.790 | 7.1 |
| | $Q_2$ | 18 | 769.523 | 699.832 | 6.7 |
| | $Q_3$ | 21 | 740.884 | 671.297 | 7.0 |
| | $Q_4$ | 24 | 774.866 | 702.969 | 7.2 |
| 2013 | $Q_1$ | 27 | 752.950 | 656.115 | 7.4 |
| | $Q_2$ | 30 | 784.344 | 689.613 | 6.7 |
| | $Q_3$ | 33 | 783.013 | 700.571 | 6.9 |
| | $Q_4$ | 36 | 854.397 | 777.185 | 6.7 |
| 2014 | $Q_1$ | 39 | 890.979 | 767.158 | 6.8 |
| | $Q_2$ | 42 | 897.411 | 789.476 | 6.0 |
| | $Q_3$ | 45 | 900.575 | 797.952 | 5.9 |
| | $Q_4$ | 48 | 939.860 | 845.044 | 5.7 |

Source: Czech Statistical Office (2017) - https://www.czso.cz/

Table 1 shows four segments therefore $k = 3$ is equal to $4 - 1$. We choose a nodal vector of dimension $+ 2 = 3 + 2 = 5$ :

$$U = (T_0, T_1, T_2, ..., T_{k+1}) = (T_0, T_1, T_2, T_3, T_4) = (0, 15, 27, 39, 50)$$

whose components create an increasing sequence. For $l = 1$ we choose from interval $< T_{l-1}, T_l > = < T_0, T_1 > = = < 0, 15 >$ points $t_{lw} = t_{1w} = 3(w - 1)$, $w = 1, 2, 3, 4, 5 = n(1)$; to number $t_{1w}$ we assign quarter $Q_{w-1}$ of year $2010 + l = 2011$, respectively quarter $Q_4$ of year 2010 (example of point $t_{11} = 0$). For example, quarter $Q_{5-1} = Q_4$ of year 2011, assigned to the number $t_{15} = 3 \cdot (5 - 1) = 12$. Furthermore, for $l = 2, 3, 4 = k + 1$ we choose from interval $< T_{l-1}, T_l >$ points $t_{lw} = 3w + 12(l - 1)$, $w = 1, 2, 3, 4 = n(l)$ ; to the number $t_{lw}$ we will assign quarter $Q_w$ of year $2010 + l$. For example, to number $t_{31} = 3 \cdot 1 + 12 \cdot (3 - 1) = 27$ we assign quarter $Q_1$ of year 2013. Or to number $t_{43} = 3 \cdot 3 + 12 \cdot (4 - 1) = 45$ we assign quarter $Q_3$ of year 2014.

**A.** We will undergo a time sequence of values $x_1$, $x_2$, $x_3$, based on table 1 to the *linear* segmented regression ($Q = 1$). A computer-generated matrix of system of normal equations (see (3)), which for $k + Q + 1 = 3 + 1 + 1 = 5$, of type 5x5, is:

$$\underline{\underline{M}} = \begin{bmatrix} 17 & 408 & 198 & 84 & 18 \\ 408 & 13464 & 7524 & 3528 & 828 \\ 198 & 7524 & 4554 & 2268 & 558 \\ 84 & 3528 & 2268 & 1260 & 342 \\ 18 & 828 & 558 & 342 & 126 \end{bmatrix},$$

and a the vector on the right side for $x_1$, respectively $x_2$, respectively $x_3$, is in given order

$$\underline{\underline{Z}}_1 = \begin{bmatrix} 13438.72 \\ 340139.19 \\ 170211.03 \\ 74841.01 \\ 16554.42 \end{bmatrix}, \quad \underline{\underline{Z}}_2 = \begin{bmatrix} 12137.84 \\ 303357.13 \\ 151155.39 \\ 66424.03 \\ 14761.54 \end{bmatrix}, \quad \underline{\underline{Z}}_3 = \begin{bmatrix} 113.8 \\ 2670.1 \\ 1271.4 \\ 519.3 \\ 104.7 \end{bmatrix}.$$

The matrix $\underline{\underline{M}}$ is symmetrical. Its element $m_{11} = 17$, which means that it is equal to the number of observed points $P_i$, $i = 1, 2, \ldots, 17$, that are in table 1. After solving the relevant system of normal equations, for example, by the Gauss' method with pivoting and normalisation, we get output parametrical equation of regressive curve (see (4)).

$$x_1 = G_1(t) = \begin{cases} 678.9213 + 6.1177 \cdot t & for \quad 0 \le t < 15 \\ 797.9583 - 1.8181 \cdot t & for \quad 15 \le t < 27 \\ 456.6459 + 10.8231 \cdot t & for \quad 27 \le t < 39 \\ 650.6281 + 5.8492 \cdot t & for \quad 39 \le t < 50 \end{cases},$$

$$x_2 = G_2(t) = \begin{cases} 651.0624 + 3.1559 \cdot t & for \quad 0 \le t < 15 \\ 736.6029 - 2.5468 \cdot t & for \quad 15 \le t < 27 \\ 432.8340 + 8.7039 \cdot t & for \quad 27 \le t < 39 \\ 505.9778 + 6.8284 \cdot t & for \quad 39 \le t < 50 \end{cases}, \qquad (5)$$

$$x_3 = G_3(t) = \begin{cases} 6.9105 - 0.0165 \cdot t & for \quad 0 \le t < 15 \\ 5.9820 + 0.0460 \cdot t & for \quad 15 \le t < 27 \\ 8.7495 - 0.0565 \cdot t & for \quad 27 \le t < 39 \\ 10.6059 - 0.1041 \cdot t & for \quad 39 \le t < 50 \end{cases}.$$

If we substitute, for example, the value of parameter $t$=35 into the equations (5) we will gain a point on the regressive curve with the following coordinates:

(835.4544, 737.4705, 6.7720).

With a small amount of prediction we could then say that in months October and November of the year 2013 export was valued at roughly 835 milliard CZK, import at roughly 737 milliard CZK and the unemployment rate roughly 6.8 percent.

The determination index for $x_1$, respectively $x_2$, respectively $x_3$, is in the following order:

$$I_{x_1}^2 = 0.9588, \quad I_{x_2}^2 = 0.9330, \quad I_{x_3}^2 = 0.7781.$$

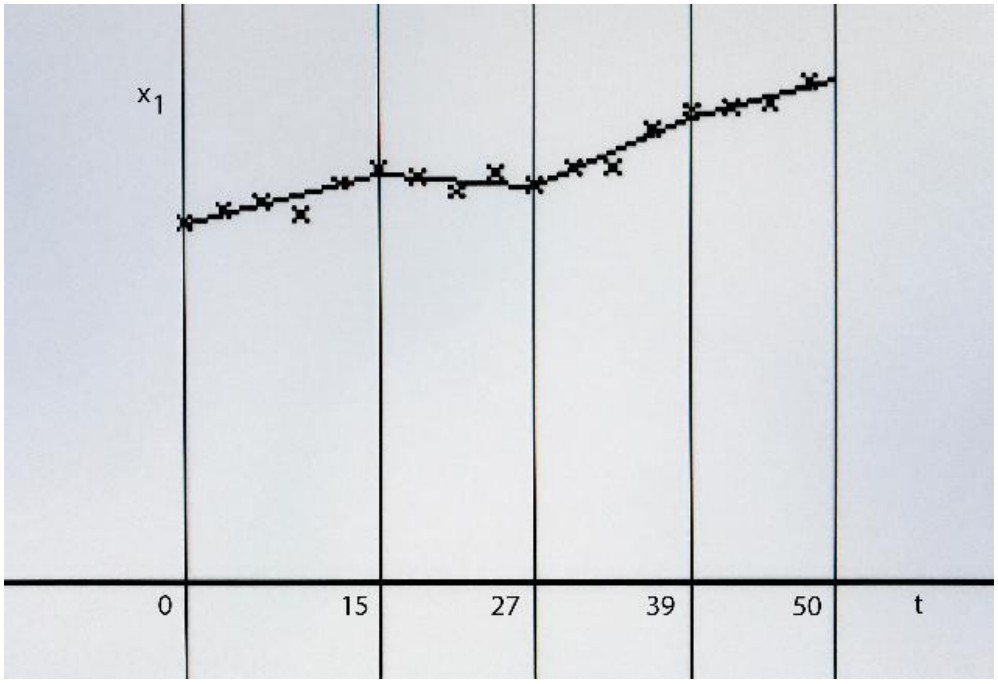
Therefore
95.88 % variability of studied values $x_1$,
93.30 % variability of studied values $x_2$,
and
77.81 % variability of studied values $x_3$,
can be explained with the chosen "linear" regression.

**B.** The chronology of the values $x_1, x_2, x_3$ based on the table 1, will be subjected to $Q = 2$ to a *quadratic* segmented regression. A computer-generated matrix $\underline{\underline{M}}$ of the system of normal equations which for $k+Q+1 = 3+2+1 = 6$ will be of type 6x6 will be symmetrical as well. The

calculated variables in this system lead to the end result to parametric equations of the output regressive curve (compare with (4)). If we were to substitute for example, the value $t = 35$ into these equations (similarly to the column **A.**) we shall get on the regressive curve a point with the coordinates

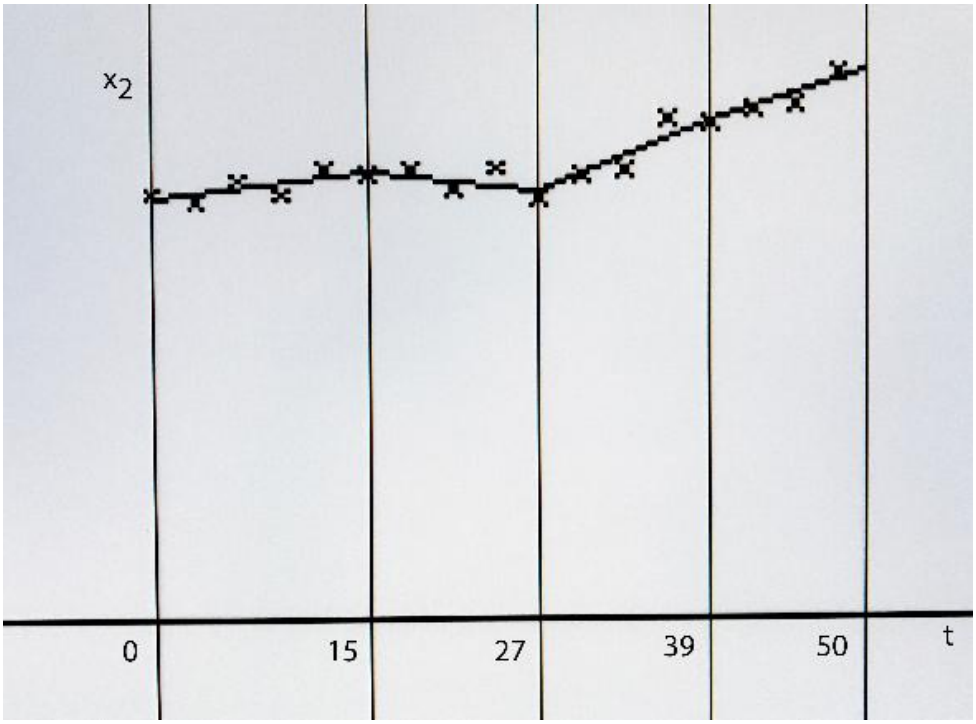
(819.4697, 726.4258, 6.7999).

Figure 1: Course of the linear regression in plane t, $x_1$



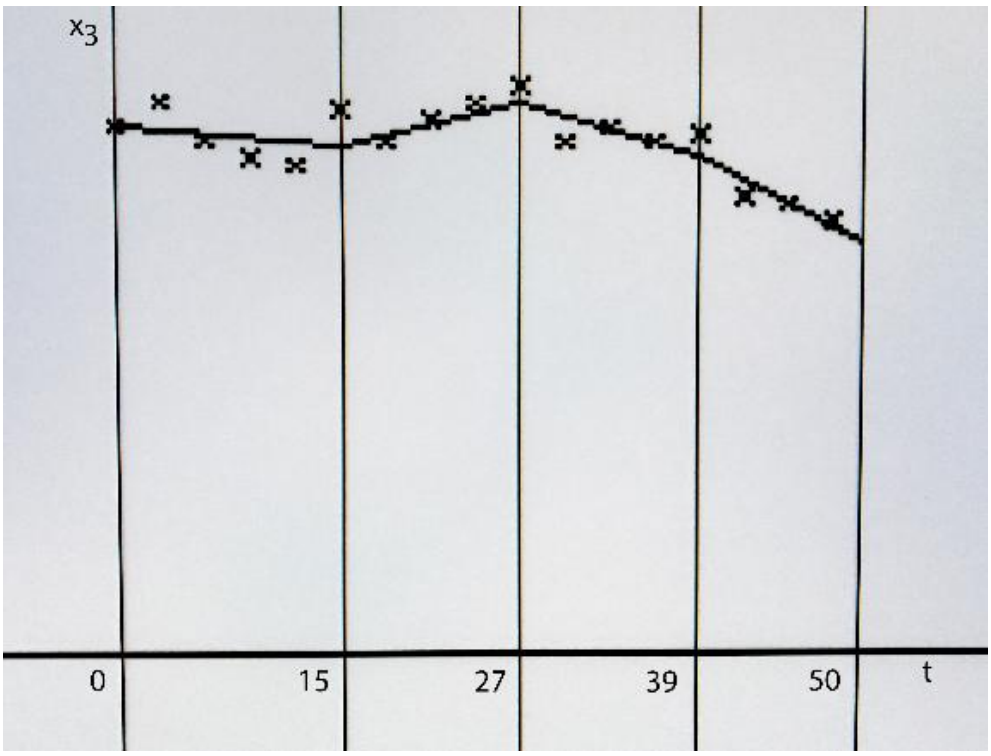

Source: Own calculation

Figure 2: Course of the linear regression in plane t, $x_2$




Source: Own calculation

Figure 3: Course of the linear regression in plane t, $x_3$




Source: Own calculation

We can then assume that in October and November of the year 2013 export was valued at roughly 819 milliard CZK, import at roughly 726 milliard CZK and the unemployment rate was roughly 6.8 %.
The determination index for $x_1$, respectively $x_2$, respectively $x_3$, is in the following order:

$$I_{x_1}^2 = 0.9388, \quad I_{x_2}^2 = 0.9015, \quad I_{x_3}^2 = 0.7858.$$

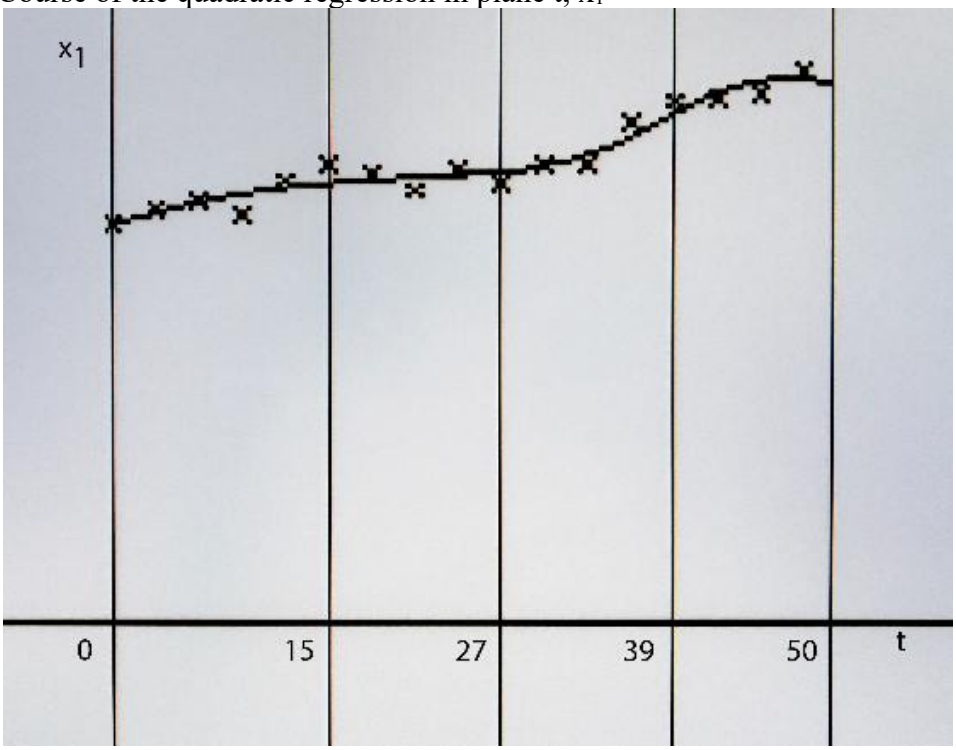
Therefore
93.88 % variability of studied values $x_1$,
90.15 % variability of studied values $x_2$,
and
78.58 % variability of studied values $x_3$,
can be explained with the chosen "quadratic" regression.
**C.** The chronology of the values $x_1, x_2, x_3$ based on the table 1, will be subjected for $Q = 3$ to a *cubic* segmented regression. A computer-generated matrix $\underline{\underline{M}}$ of the system of normal equations which for $k+Q+1 = 3+3+1 = 7$ will be of type 7x7 will be symmetrical as well. The calculated variables in this system lead to the end result to parametric equations of the output regressive curve (compare with (4)). If we were to substitute, for example, the value $t = 35$ into these equations, we shall get on the regressive curve a point with the coordinates

$$(835.9608, 741.9907, 6.8105).$$

We can then assume that in October and November of the year 2013 export was valued at roughly 836 milliard CZK, import at roughly 742 milliard CZK and the unemployment rate was roughly 6.8 %.
The determination index for $x_1$, respectively $x_2$, respectively $x_3$, is in the following order:

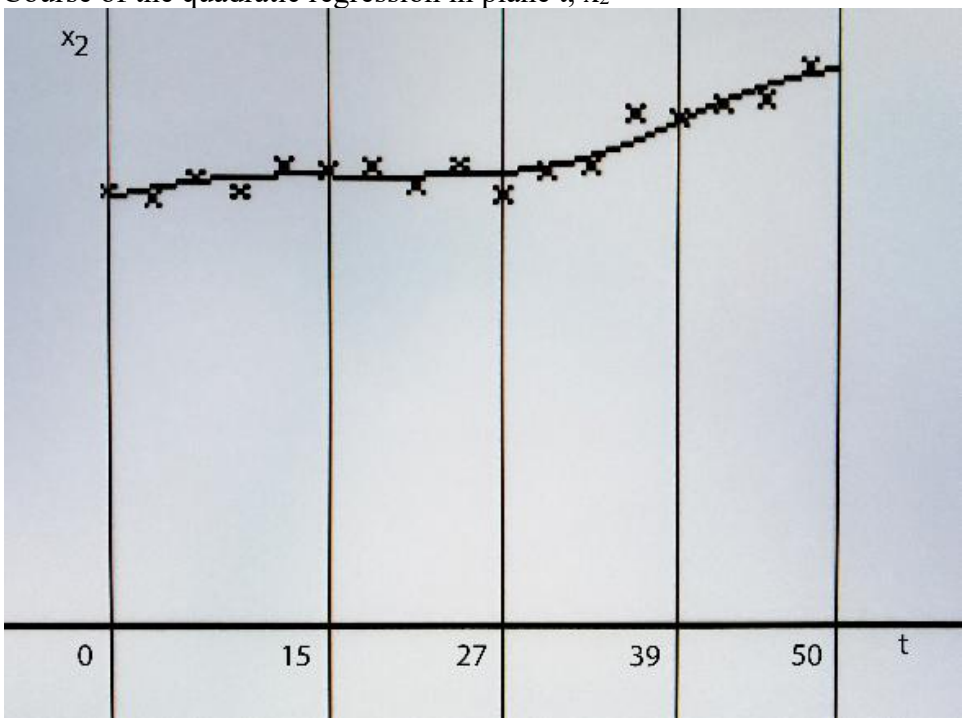
$$I_{x_1}^2 = 0.9587, \quad I_{x_2}^2 = 0.9408, \quad I_{x_3}^2 = 0.7849.$$

Figure 4: Course of the quadratic regression in plane t, $x_1$




Source: Own calculation

Figure 5: Course of the quadratic regression in plane t, $x_2$

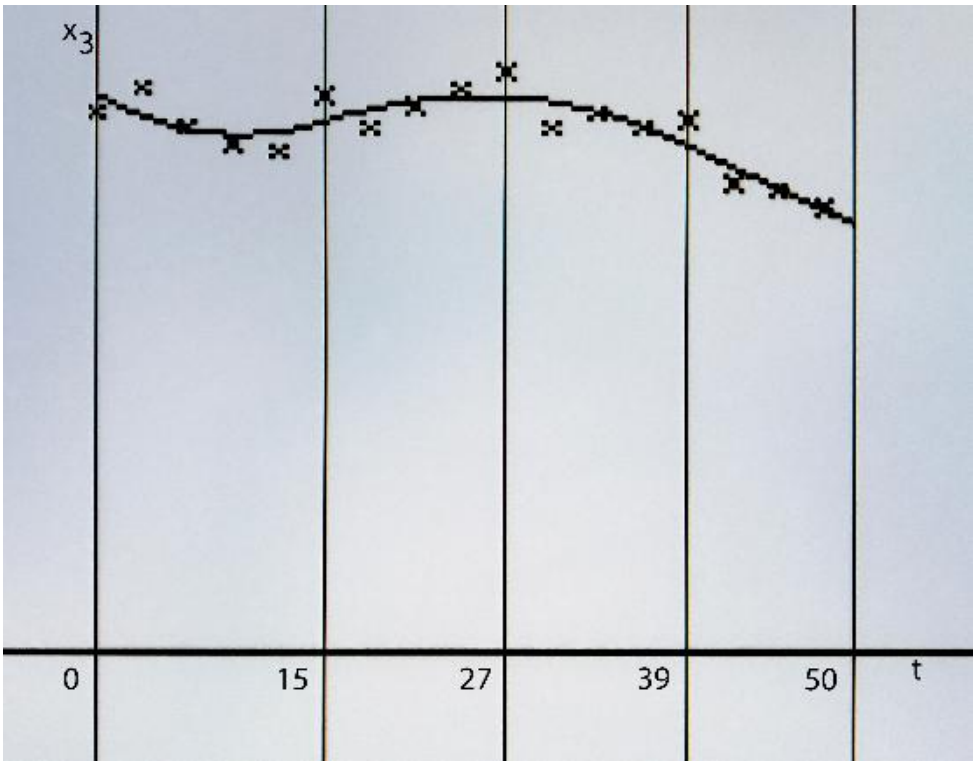



Source: Own calculation

Therefore
95.87 % variability of studied values $x_1$,
94.08 % variability of studied values $x_2$,
and
78.49 % variability of studied values $x_3$,

can be explained with the chosen "cubic" regression.

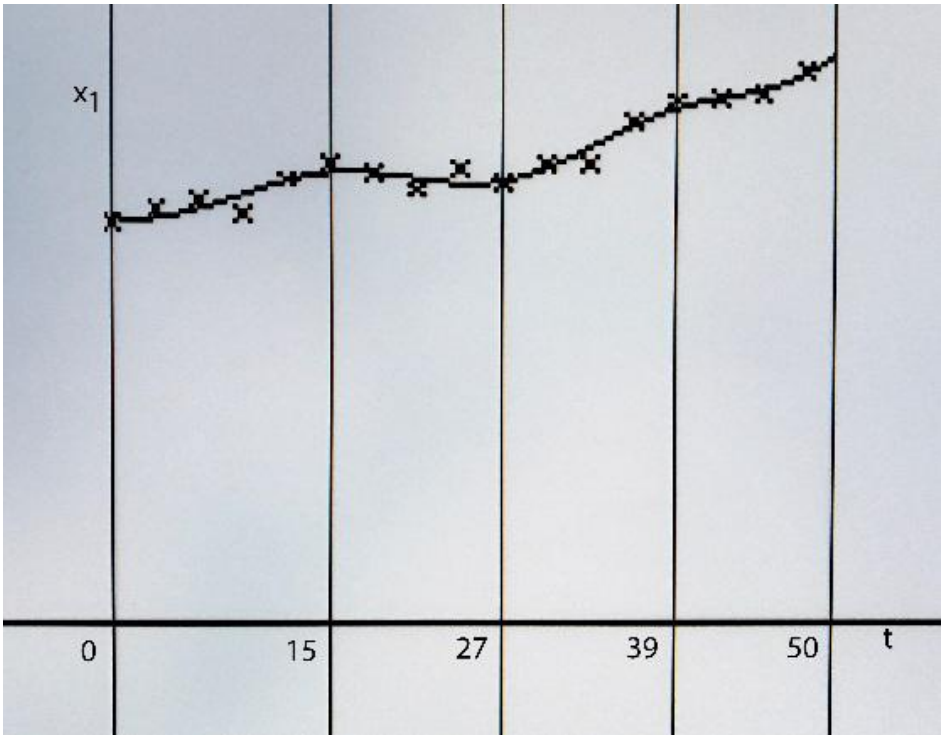
Figure 6: Course of the quadratic regression in plane t, $x_3$




Source: Own calculation

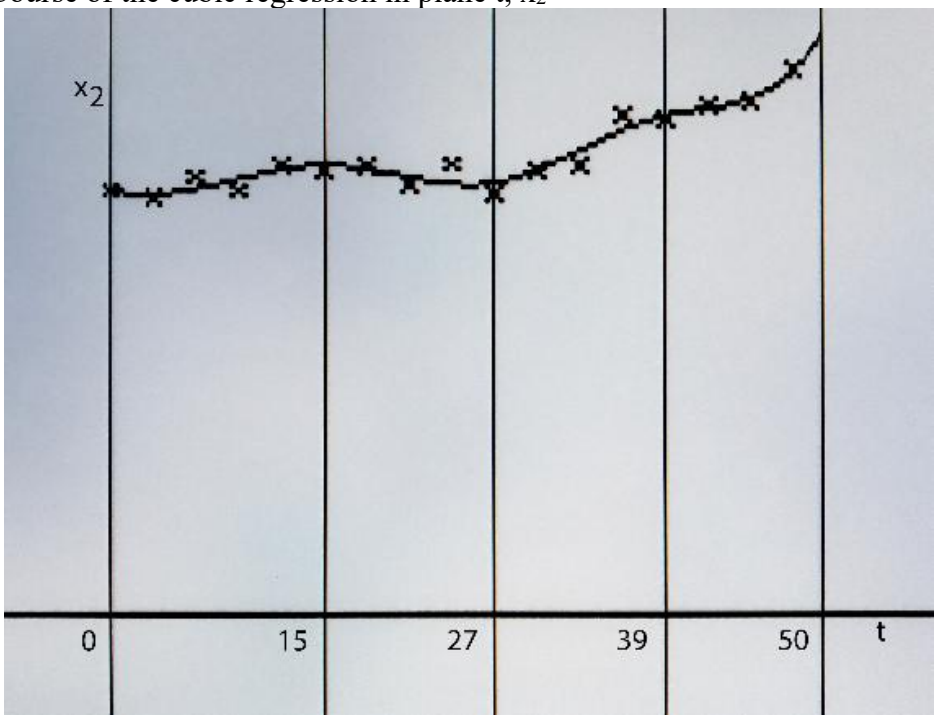
## 3. Remark

Figure 7: Course of the cubic regression in plane t, $x_1$



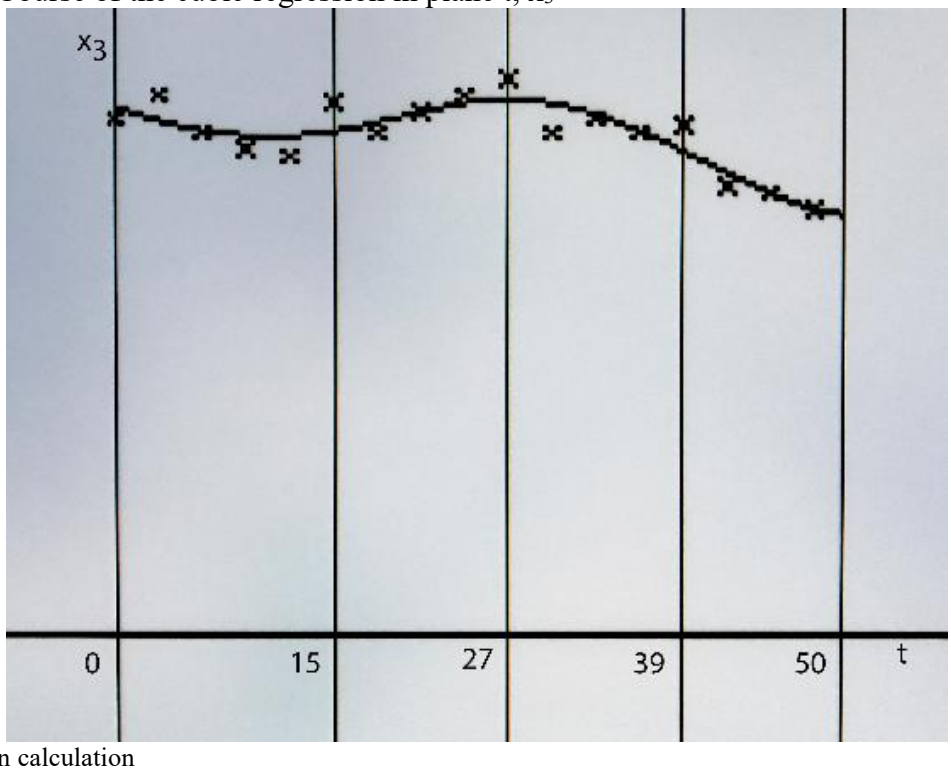

Source: Own calculation

Figure 8: Course of the cubic regression in plane t, $x_2$




Source: Own calculation

Figure 9: Course of the cubic regression in plane t, $x_3$




Source: Own calculation

As I mentioned in the beginning, the segmented regression can also be based on something else than the cut-off polynomials, for example on so called B-spline basis functions. You can read about this in detail, in a paper by Kaňka, M. (2015, pp. 47-66). Also for getting to know more about spline functions, the author recommends two essential sources, that you can find in references, which are, Bézier, P. (1972), and Böhmer, K. (1974).

Because there is certain connection between cut-off polynomials and B-spline basis functions, the author of this paper thought it would be good to reference on this paper, which has been written by Boor, C. D. (1972, pp. 50-62), in which de Boor derives an important recursive formula for the calculation of B-spline basis functions.
The introduction of B-spline basis functions is based on a theory of so called variable differences of real functions, which has its place in an area called interpolation. The author thought it would be a good idea to mention this too, see Schrutka, L. (1945, pp. 2-31). The last two references are aimed at the practical use of segmented regression in economic questions, those are Guzik, B. (1974, pp. 11-27) and Feder, P. I. (1975).

## 4. Conclusion

The author, who wrote the program for *linear* $(Q = 1)$, *quadratic* $(Q = 2)$ and *cubic* $(Q = 3)$ segmented regression, ranked the individual phases the following way:

a) calculation of the matrix and vectors of the right sides of the system of normal equations of the studied variables for linear ($Q$=1), quadratic ($Q$=2) and cubic ($Q$=3) models;
b) calculations of unknowns in a system of normal equations using the smooth function with pivoting and normalization;
c) calculation of parametric equations of output regressive curve in the chosen model;
d) calculation of indexes of determination of the observed variables in the chosen model;
e) calculation of coordinates of a point on the output regressive curve for the chosen parameter value.

The program with instructions for use can be obtained by e-mail on request from the author of this paper.
The author would like to thank prof. Ing. R. Hindls, CSc. who helped to obtain the values of the observed variables mentioned in the table 1. This article has been written thanks to his inspiration.

## References

[1] Bézier, P. 1972. Numerical control; mathematics and applications. London: J. Wiley. ISBN-13: 9780471071952, ISBN-10: 0471071951.

[2] Boor, C. D. 1972. On calculating with B-splines. Journal of Approximation Theory, vol. 6, iss. 1, pp. 50-62. doi:10.1016/0021-9045(72)90080-9.

[3] Böhmer, K. 1974. Spline Funktionen. Stuttgart: B. G. Teubner. ISBN-13: 9783519020479, ISBN-10: 3519020475.

[4] Kaňka, M. 2015. Segmented Regression Based on B-splines with solved Examples. Statistics and Economy Journal, vol. 95, iss. 4, pp. 47-66.

[5] Kaňka, M. 2016. Segmented Regression Based on Cut-off polynomials. Statistics and Economy Journal, vol. 96, iss. 2, pp. 60-72.

[6] Makarov, V. L., Chlobystov, V. V. 1983. Splain-approximaciia funkcii. Moscow: Vvssaia skola. Pp. 4-35.

[7] Meloun, M., Militky, J. 1994. Statistical Treatment of Experimental Data (in Czech language). Prague: Edice Plus. Pp. 752-759. ISBN: 80-85297-56-6

[8] Guzik, B. 1974. Estymatory horyzontu prognozy w trendach segmentowych. Przeglad Statystyczny, vol. 21, pp. 11-27.

[9] Feder, P. I. 1975. On Asymptotic Distribution Theory in Segmented Regression Problems. The Annals of Statistics, vol. 3, iss. 1, pp. 49-83. doi:10.1214/aos/1176342999.

[10] Seger, J. 1988. Statisticke metody pro ekonomy prumyslu, Statistical methods for industrial economists. Prague: Statni nakladatelstvi technicke literatury Alfa. Pp. 431-436.

[11] Schrutka, L. 1945. Leitfaden der Interpolation. Wien: Springer-Verlag. Pp. 2-31.