ON SOME STATISTICAL MODELS FOR ASSESSING THE EVOLUTION OF SALES OF CERTAIN PRODUCTS

CĂTĂLIN ILIE MITRAN [*]

ABSTRACT: *The mathematical statistic models are very used in solving of many types of engineering, economic and other practical issues. This paper aims to analyse the evolution of the prices of a particular type of a product for one year. For this purpose, will be used the smallest squares method which in this case is equivalent to the linear regression method and also some typical mathematical statistics methods. The effectiveness of the proposed methods should be tested by verifying some practical hypothesis such as those of the linear regression method.*

KEY WORDS: *Linear regression method, econometric model, three sigma rule, residual variable.*

JEL CLASSIFICATION: *C40.*

1. INTRODUCTION

Let us suppose that we have registered a lot of values consisting in the amounts necessary to obtain a particular type of product on the one hand and the proceeds from the sale of that product on the other hand. We suppose the values were recorded for one year. In other words, we have the next data:

Month	x (RON)	y (RON)
1	4200	8450
2	4800	9450
3	5000	10100
4	4500	9000
5	4700	9400
6	5500	11000
7	6000	11700
8	5400	10650

[*] *Lecturer, Ph.D., University of Petroşani, Romania*

9	5000	9800
10	4500	8900
11	4300	8600
12	5300	10450

What we propose to find is a link between the two set of values, which, of course, can be described in terms of mathematics and for which other mathematical methods can be found for analysing and interpreting the results. First, for data analysis we will choose the appropriate econometric model whose parameters can be estimated. After that, we will verify the model assumption chosen on the basis of a specific working methods, in this case based on some statistical-mathematical methods.

First of all, let us remind some notions in mathematical statistics that will be used further. Let us consider a discrete random variable:

$$X: \begin{pmatrix} x_1 & x_2 & \dots & x_n \\ p_1 & p_2 & \dots & p_n \end{pmatrix}$$

where $x_1, x_2, \dots, x_n$ are the elementary events of the experiment described by the random variable and $p_1, p_2, \dots, p_n$ the corresponding probabilities having the properties:

1) $p_i \geq 0, \forall i = 1, \dots, n$;
2) $\sum_{i=1}^{n} p_i = 1$.

Let us remember that we have for the average value, dispersion, average square deviation and the correlation coefficient the next formulas:

$$m = M(X) = \sum_{i=1}^{n} x_i p_i \ ,$$

$$D^2(X) = \sum_{i=1}^{n} (x_i - m)^2 p_i \ ,$$

$$\sigma_x = \sqrt{D^2(X)} \ ,$$

$$r_{X,Y} = \frac{\sum_{i=1}^{12} (x_i - M(X))(y_i - M(Y))}{n \sigma_x \sigma_y} \ .$$

2. DATA ANALYSIS

During the data analysis process will be considered the choice of the econometric model describing the link between variables, the estimation of the linear regression parameters and the correlation coefficient.

2.1. The choice of the econometric model describing the link between the variables

Taking care of the distribution of the pairs $(x_i, y_i), i = 1, \dots, 12$ it is easily to observe that all of these are situated around of a line. The algebric description of this line is given by the equation:

$$y(x) = ax + b \tag{1}$$

and the link between the input data is of the next form:

$$y_i = ax_i + b + \varepsilon_i, i = 1, \dots, 12 \tag{2}$$

2.2. The estimation of the linear regression model parameters

The minimization criteria for the approximation errors using the least squares method is given by:

$$\sum_{i=1}^{12} \varepsilon_i^2 = \sum_{i=1}^{12} [y_i - (ax_i + b)]^2 \to min. \tag{3}$$

Denoting by $F(a,b) = \sum_{i=1}^{12} [y_i - (ax_i + b)]^2$ and imposing the singular point conditions, that means

$$\begin{cases} \frac{\partial F(a,b)}{\partial a} = 0 \\ \frac{\partial F(a,b)}{\partial b} = 0 \end{cases}. \tag{4}$$

Making calculations will result a linear system of unknowns *a* and *b*:

$$\begin{cases} a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a \sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i \end{cases} \tag{5}$$

having the system determinant

$$\Delta = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} > 0$$

which implies a single solution which can be determined using the Cramer's rule:

$$a^* = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & n \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix}},$$

$$b^* = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i \end{vmatrix}}{\begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix}}.$$

Next, calculating the values:

$$\Delta_1 = \frac{\partial^2 F}{\partial a^2}(a^*, b^*) = \sum_{i=1}^{n} x_i^2 > 0,$$

$$\Delta_2 = \begin{vmatrix} \frac{\partial^2 F}{\partial a^2}(a^*, b^*) & \frac{\partial^2 F}{\partial a \partial b}(a^*, b^*) \\ \frac{\partial^2 F}{\partial b \partial a}(a^*, b^*) & \frac{\partial^2 F}{\partial b^2}(a^*, b^*) \end{vmatrix} = \begin{vmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{vmatrix} = \Delta > 0,$$

using the properties of the extreme functions we deduce that (a^*, b^*) is a minimum point.

Let us consider the following table of data:

Coefficient	x_i	y_i	x_i^2	$x_i y_i$
1	4200	8450	17.640.000	35.490.000
2	4800	9450	23.040.000	45.360.000
3	5000	10100	25.000.000	50.500.000
4	4500	9000	20.250.000	40.500.000
5	4700	9400	22.090.000	44.180.000
6	5500	11000	30.250.000	60.500.000
7	6000	11700	36.000.000	70.200.000
8	5400	10650	29.160.000	57.150.000
9	5000	9800	25.000.000	49.000.000
10	4500	8900	20.250.000	40.050.000
11	4300	8600	18.490.000	36.980.000
12	5300	10450	28.090.000	55.385.000
Sum	59200	117500	295.260.000	585.295.000

The system becomes:

$$\begin{cases} 295.260.000a + 59200b = 585.295.000 \\ 59200a + 12b = 117.500 \end{cases}$$

having the solution $a \approx 1{,}7552, b \approx 1132{,}7702$

We have the approximation equation

$$y(x) = ax + b = 1{,}7552a + 1132{,}7702 \tag{6}$$

This kind of approximation allow us to make certain predictions for the next time period. So, assuming that the next month we have the amounts necessary for making the product equal with 5000 unitary units we will get an approximation of

$$y(5000) = 5000a + b = 1{,}7552 \cdot 5000 + 1132{,}7702 \simeq 9909$$

monetary units to be cashed.

We also to calculate the values

$$\hat{y}_i = ax_i + b, i = 1, \dots, 12 \tag{7}$$

which represents the approximation values for $y_i, i = 1, \dots, n$ using the approximation equation.

We get the values:

i	$\hat{y}_i$
1	8504,61
2	9577,73
3	9908,77
4	9031,17
5	9382,21
6	10786,37
7	11663,97
8	10610,85
9	9908,77
10	9031,17
11	8680,13
12	10435,33

2.3. The correlation coefficient

First of all, let us see that our data can be organized like two discrete random variables:

$$X: \begin{pmatrix} 4200 & 4800 & \dots & 4300 & 5300 \\ \frac{1}{12} & \frac{1}{12} & \dots & \frac{1}{12} & \frac{1}{12} \end{pmatrix}, \ Y: \begin{pmatrix} 8450 & 9450 & \dots & 8600 & 10450 \\ \frac{1}{12} & \frac{1}{12} & \dots & \frac{1}{12} & \frac{1}{12} \end{pmatrix}$$

We have to calculate the correlation coefficient:

$$r_{X,Y} = \frac{\sum_{i=1}^{12}(x_i - M(X))(y_i - M(Y))}{n\sigma_x\sigma_y} \tag{8}$$

We get successively the next values:

$$M(X) = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{59200}{12} \approx 4933,33 \, , M(Y) = \frac{\sum_{i=1}^{12} y_i}{12} = 9791,66$$

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{12}(x_i - M(X))^2}{12}} \approx 518,01$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{12}(y_i - M(Y))^2}{12}} \approx 969,50$$

and taking care of all these values we find for the correlation coefficient the next value: $r_{X,Y} = 1{,}075$, a value very close to 1, which means that there is a very strong linear dependence between the two regression variables.

3. VERIFYING THE HYPOTHESIS OF THE LINEAR REGRESSION MODEL

Two hypotheses considered to be the most important in estimated and establishing the linear regression model will be considered. It is the hypothesis in which the series of data are not affected by measurement errors and the one in which the residual variable is of average null.

In the first case, the hypothesis that the series are not affected by measurement errors is done by checking the condition

$$x \in (M(X) \pm 3\sigma_x), y \in \left(M(Y) \pm 3\sigma_y\right) \tag{9}$$

and in the second case, the hypothesis that the residual variable is of average null, by checking the contion

$$M(\varepsilon) = \frac{\sum_{i=1}^{1n} \varepsilon_i}{n} \simeq 0. \tag{10}$$

3.1. The series are not affected by measurements errors

To verify this hypothesis, we will apply the so-called" three sigma rule", that says that the variables *x* and *y* must verify the relationships (9):

$$x\epsilon(M(X) \pm 3\sigma_x), y\epsilon\left(M(Y) \pm 3\sigma_y\right).$$

The average squares deviations σ_x and σ_y are, in this case, determined from the relationships:

$$\sigma_x = \sqrt{\frac{\sum_{i=1}^{12}\left(x_i - M(X)\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{12}\left(x_i - M(X)\right)^2}{12-1}} = \sqrt{\frac{\sum_{i=1}^{12}\left(x_i - M(X)\right)^2}{11}} \simeq$$
$$\simeq 541{,}04$$

$$\sigma_y = \sqrt{\frac{\sum_{i=1}^{12}\left(y - M(Y)\right)^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{12}\left(y_i - M(Y)\right)^2}{12-1}} = \sqrt{\frac{\sum_{i=1}^{12}\left(y_i - M(Y)\right)^2}{11}} \simeq$$
$$\simeq 1012{,}61$$

In this case, relationship (9) becomes equivalent to:

$$\begin{cases} 4933{,}33 - 3 \cdot 541{,}04 < x_i < 4933{,}33 + 3 \cdot 541{,}04, \forall i = 1, \dots, 12 \\ 9791{,}66 - 3 \cdot 1012{,}61 < y_i < 9791{,}66 + 3 \cdot 1012{,}61, \forall i = 1, \dots, 12 \end{cases}$$

that means

$$3310{,}15 < x_i < 6556{,}45, \forall i = 1, \dots, 12$$
$$6753{,}83 < y_i < 12835{,}49, \forall i = 1, \dots, 12$$

both inequalities being true because the values $x_i, i = 1, \dots, 12$ oscillates between minimum value 4200 and the maximum value 6000 meanwhile the values $y_i, i = 1, \dots, 12$ oscillates between theminimum value 8450 and the maximum value 11700.

3.2. The residual variable has an average value of 0 (average null)

Starting from the approximation relationship:

$$y_i = ax_i + b + \varepsilon_i, \forall i = 1, \dots, 12 \Leftrightarrow y_i = \hat{y}_i + \varepsilon_i, \forall i = 1, \dots, 12 \Leftrightarrow$$
$$\Leftrightarrow \varepsilon_i = y_i - \hat{y}_i, \forall i = 1, \dots, 12$$

we find the following values for errors:

i	$\hat{\varepsilon}_i$
1	-54,53
2	-107,65
3	191,31
4	-31,09
5	17,87
6	213,71
7	36,11
8	39,23
9	-108,69
10	-131,09
11	-80,05
12	14,75

The average value for these errors is given by:

$$M(\varepsilon) = \frac{\sum_{i=1}^{12} \varepsilon_i}{12} = \frac{-0{,}12}{12} = -0{,}01$$

a value very close to 0, which verifies the hypothesis of the average value being 0.

4. CONCLUSIONS

The least squares method and, in particular, the linear case of this method usually called linear regression is a very used method in many types problems of approximation as was the case in the considered economic example. The degree of safety of the linear regression model is given to very good extent by verifying the hypothesis proposed and proving to be a high one.

REFERENCES:

[1]. Despa, R.; Vişan, C. (2003) *Mathematics applied in economics*, Sylvi Publishing House, Bucharest

[2]. Mitran, I. (2009) *Modelling of technical and economic decision*, AGIR Publishing House, Bucharest

[3]. Popescu, O. (1993) *Mathematics applied in economics*, vol. I, II, Didactic and Pedagogical Publishing House, Bucharest

[4]. Purcaru, I.; Berbec, F.; Dan, S. (1996) *Financial mathematics*, Economic Publishing House, Bucharest

[5]. Stroe, R. (2000) *Modelling financial decisions*, Academy of Economic Science Publishing House, Bucharest