TJMM
15 (2023), No. 1-2, 103-113

APPLICATION OF THE LEAST SQUARES METHOD TO STUDY THE CORRELATION BETWEEN CLIMATIC FACTORS IN ROMANIA

ANAMARIA POPESCU

Abstract. The paper presents linear and non-linear regression mathematical models that estimate the evolution of processes or phenomena based on some parameters that define the processes and phenomena in order to perform calculations and approximations of experimental data. Having a series of data on climatic factors, the analysis of the results from a period from 1901 to the present, consisting of decades, is presented, approaching the approximation by the least squares method and verifying the existence of a correlation and identifying the model.

1. Introduction

The method of least squares is a mathematical method of obtaining a solution of a system of overdetermined equations, i.e. having more equations than unknowns. Least squares means that the solution obtained minimizes the sum of the squares of the deviations from the values of the equations.

The most important application is to determine the coefficients of a mathematical function that best approximates a set of data. This best approximation minimizes the squares of the deviations between the given values and those calculated using that function.

There are two variants of the least squares method:

- The linear least squares method, which solves problems based on systems of linear equations. An example of such an application is linear regression, which is widely used in statistics and experimental data processing. Solving the resulting system of equations is usually done by direct methods.
- The nonlinear least squares method, which solves problems based on systems of nonlinear equations. Solving the resulting system of equations is usually done by iterative methods, with each iteration using linearization.

The method was first developed by Carl Friedrich Gauss around 1794 [3, 8].

In the following, we will refer to the situation of linear regression (the relationship between the two variables can be described by a line within the point cloud), parabolic and cubic regression. Regression is closely related to the concept of correlation. If we had a perfect correlation, the estimation would be extremely accurate.

Climate warming is a phenomenon unanimously accepted by the international scientific community, and is already evident from the analysis of observational data over long periods of time. Simulations with complex global climate models have shown that the main drivers are both natural (variations in solar radiation and volcanic activity) and anthropogenic (changes in atmospheric composition due to human activities). Only the cumulative effect of these two factors can explain the observed changes in global average air and ocean temperature, melting snow and ice and rising global average sea level.

2010 *Mathematics Subject Classification.* 62-07, 62-09, 62G05, 62J02, 62J05, 62M10, 91B84.
Key words and phrases. Least squares method, regression equation, correlation.

The increase in the concentration of greenhouse gases in the atmosphere, especially carbon dioxide, has been the main cause of the pronounced warming of the last 50 years of the 20th century (0.13° C/decade), and is about double the value of the last 100 years (0.74° C over the period 1906-2005), as shown in the Fourth Assessment Report of the Intergovernmental Panel on Climate Change [8].

The aim of the paper was to present the annual amount of precipitation as a function of the average annual air temperature. Statistical data from the Statistical Yearbook of Romania were used for this study.

2. Approximation by the least squares method

We consider the real function $f : [a, b] \to \mathbb{R}$, for which the values $y_i = f(x_i)$ are known in $n + 1$ distinct points x_i, $i = \overline{0, n}$, from the interval $[a, b]$, that is, the value pairs:

$$(x_0, y_0),\ (x_1, y_1),\ (x_2, y_2), ..., (x_n, y_n). \tag{1}$$

In the general case, the points can be any, but they are usually equidistant with the discretization step h :

$$x_0 = a,\ x_n = b,\ x_{i+1} - x_i = h = \frac{b-a}{n},\ i = \overline{0, n-1}. \tag{2}$$

It is required to determine the polynomial $P_m,\ grad\ P_m = m < n$, of the form:

$$P_m(x) = c_0 + c_1x_1 + c_2x_2 + \cdots + c_mx_m,\ \forall x \in [a, b], \tag{3}$$

which approximates the function so that the sum of the squares of the differences between the approximate and exact values at the $(n + 1)$ points is minimized. In other words, the following optimization problem must be solved:

$$\hat{P}_m = \left\{ P_m |\min_{c_0,...,c_n} F,\ F = \sum_{i=0}^{n} \left[P_m(x_i) - y_i\right]^2 \right\} \tag{4}$$

The resulting calculation method is called the least squares method (LSM) and is used when either the pairs (1) are not known exactly or is very large.

Approximating the function known in the form of the set of values (1) by a polynomial of the form (3) by LSM is generally also called polynomial regression, with the widely used specializations linear regression ($m = 1$), parabolic regression ($m = 2$) and cubic regression ($m = 3$).

2.1. **Linear polynomial approximation (m=1) by LSM.** The LSM algorithm ([2]) is based on the condition that the sum of the squares of the differences Δy_i is minimal, where:

$$\Delta y_i = y_i - f(c_0, c_1, x_i) \tag{5}$$

$$S = \sum_i (\Delta y_i)^2 = min. \tag{6}$$

The index of each sum takes integer values in the interval $[1, n]$, n = the number of values x_i, respectively y_i. This method will be applied after testing the level of errors and eliminating gross errors.

The functional dependence is searched in the form $y = c_0 + c_1x$.

For the calculation of c_0 and c_1 we have the following system of equations:

$$\begin{cases} nc_0 + c_1 \sum_i x_i = \sum_i y_i \\ c_0 \sum_i x_i + c_1 \sum_i x_i^2 = \sum_i x_i y_i \end{cases} \tag{7}$$

where:

n is the number of investigated cases;

y is the estimated result;

c_0 is the intercept (the place on the ordinate where the regression line intersects with OY, the value of y for $x = 0$);

c_1 is the regression slope (it tells us how much y changes when x increases (decreases) by one unit;

x is the (known) criterion variable.

The calculation of the regression coefficients c_0 and c_1, respectively, leads to the realization of the first step in the regression process. By means of regression, predictions can be made of one variable, depending on the value of another. Prediction is the process of estimating the value of one variable knowing the value of another variable.

2.2. **Parabolic polynomial approximation (m=2) by LSM.** The parabolic approximation polynomial function, also called parabolic regression, is sought in the form $y = c_0 + c_1 x + c_2 x^2$.

The function F to be minimized, viewed as a function of the variables $c_0,\ c_1,\ and\ c_2$:

$$F = \sum_{i=0}^{n} \left(c_0 + c_1 x + c_2 x^2 - y_i\right)^2. \tag{8}$$

To minimize the convex function F, it is sufficient to cancel its partial derivatives, thus obtaining the following linear system of equations:

$$\begin{cases} (n+1)\ \mathrm{c}_0 + c_1 \sum_i x_i + c_2 \sum_i x_i^2 = \sum_i y_i \\ c_0 \sum_i x_i + c_1 \sum_i x_i^2 + c_2 \sum_i x_i^3 = \sum_i x_i y_i \\ c_0 \sum_i x_i^2 + c_1 \sum_i x_i^3 + c_2 \sum_i x_i^4 = \sum_i x_i^2 y_i \end{cases} \tag{9}$$

2.3. **Cubic polynomial approximation (m=3) by LSM.** The cubic approximation polynomial function, also called cubic regression, is sought in the form $y = c_0 + c_1 x + c_2 x^2 + c_3 x^3$.

The function F to be minimized, viewed as a function of the variables $c_0,\ c_1,\ c_2\ and\ c_3$:

$$F = \sum_{i=0}^{n} \left(c_0 + c_1 x + c_2 x^2 + c_3 x^3 - y_i\right)^2 \tag{10}$$

As in the case of the parabolic polynomial approximation, the following linear system of equations is required to determine the coefficients $c_0, c_1, c_2\ and\ c_3$:

$$\begin{cases} (n+1)\ \mathrm{c}_0 + c_1 \sum_i x_i + c_2 \sum_i x_i^2 + c_3 \sum_i x_i^3 = \sum_i y_i \\ c_0 \sum_i x_i + c_1 \sum_i x_i^2 + c_2 \sum_i x_i^3 + c_3 \sum_i x_i^4 = \sum_i x_i y_i \\ c_0 \sum_i x_i^2 + c_1 \sum_i x_i^3 + c_2 \sum_i x_i^4 + c_3 \sum_i x_i^5 = \sum_i x_i^2 y_i \\ c_0 \sum_i x_i^3 + c_1 \sum_i x_i^4 + c_2 \sum_i x_i^5 + c_3 \sum_i x_i^6 = \sum_i x_i^3 y_i \end{cases} \tag{11}$$

3. The least squares method for studying the correlation between the average annual air temperature and the annual amount of precipitation

In this paragraph, we will study the correlation between the annual amount of precipitation and the average annual air temperature. We include in the table below the data taken from the Statistical Yearbook of Romania, 2022 edition [11, 12]: X = Average annual air temperature, Y = annual amount of precipitation, mm.

YEAR	X	Y
1901	10,1	589,6
1910	10,65	461,5
1920	9,9	476,5
1930	11	533,4
1940	8,45	645,4
1950	11,1	497,8
1960	10,85	635,3
1970	10,3	719
1980	9,2	726
1990	11,25	466,3
2000	11,65	386,7
2010	11,25	802,6
2020	12,4	589,9
2021	11,45	638,8
2022	11,77	528,9

Table 1. The evolution of the average annual air temperature and the annual amount of precipitation in Romania during the period 1901-2022

$n = 15$ $\quad \sum x^3 = 19126,33911$ $\quad \sum xy = 93046,153$

$\sum x = 161,32$ $\quad \sum x^4 = 210552,3868$ $\quad \sum x^2y = 1004466,709$

$\sum y = 8697,7$ $\quad \sum x^5 = 2332899,874$ $\quad \sum x^3y = 10934063,21$

$\sum x^2 = 1749,7504$ $\quad \sum x^6 = 26001910,65$

3.1. **Linear polynomial least squares method (m=1).** The functional dependence is searched in the form $y = c_0 + c_1x$.

System (7) will be written:

$$\begin{cases} 15\ \mathrm{c}_0 + 161,32c_1 = 8697,7 \\ 161,32c_0 + 1749,75c_1 = 93046,15 \end{cases}$$

The solution of the system is: $c_0 = 939,153$ and $c_1 = -33,4093$.

The obtained regression equation is:

$$y = 939,153 - 33,4093\ x. \tag{12}$$

We have the following predictions: for the average annual air temperature of 11 degrees, we will have the annual amount of precipitation of 571,7 mm.

3.2. **Parabolic polynomial least squares method (m=2).** The parabolic approximation polynomial function, also called parabolic regression, is sought in the form $y = c_0 + c_1x + c_2x^2$.

System (9) will be written:

$$\begin{cases} 16\ \mathrm{c}_0 + 161,32c_1 + 1749,75c_2 = 8697,7 \\ 161,32c_0 + 1749,75c_1 + 19126,33911c_2 = 93046,153 \\ 1749,75c_0 + 19126,33911c_1 + 210552,3868c_2 = 1004466,709 \end{cases}$$

The solution of the system is: $c_0 = 1800,2$, $c_1 = -200,51$ and $c_2 = 8,024$.

Parabolic regression is:

$$y = 1800,2 - 200,51\ x + 8,024\ x^2. \tag{13}$$

Accordingly, we can make the following predictions: for the average annual air temperature of 11 degrees, we will have the annual amount of precipitation of 565,5 mm.

3.3. **Cubic plynomial least squares method (m=3).** The cubic approximation polynomial function (cubic regression) has the form $y = c_0 + c_1x + c_2x^2 + c_3x^3$.

System (11) will be written:

$$\begin{cases} 16\ \mathrm{c}_0 + 161,32c_1 + 1749,75c_2 + 19126,33911c_3 = 8697,7 \\ 161,32c_0 + 1749,75c_1 + 19126,33911c_2 + 210552,3868c_3 = 93046,153 \\ 1749,75c_0 + 19126,33911c_1 + 210552,3868c_2 + 2332899,874c_3 = 1004466,709 \\ 19126,33911c_0 + 210552,3868c_1 + 2332899,874c_2 + 26001910,65c_3 = 10934063,21 \end{cases}$$

The solution of the system is: $c_0 = -6772,5$, $c_1 = 2301,5$, $c_2 = -233,26$ and $c_3 = 7,6937$.

Cubic regression is:

$$y = -6772,5 + 2301,5\ x - 233,26\ x^2 + 7,6937\ x3. \tag{14}$$

Accordingly, we can make the following predictions: for the average annual air temperature of 11 degrees, we will have the annual amount of precipitation of 559,9 mm.

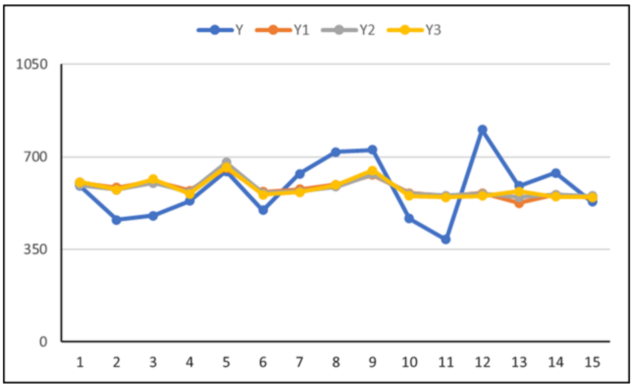


FIGURE 1. Graphical representations of statistical data (Y), linear regression (Y1), parabolic regression (Y2) and cubic regression (Y3)

4. Verification of the existence of a correlation and identification of the model

In order to be able to properly analyze the existing correlation between the two climatic indicators presented in the previous table, it is necessary that in a first step of this research we identify a series of particularities aimed at the evolution of each quantity considered in the time interval under analysis. In this sense, with the help of the computer program Eviews 12 we studied, in a first stage, the individual evolution of the two indicators [6].

Thus, studying the evolution of the average annual air temperature in Romania during the period 1901-2022 allowed obtaining the following information and significant graphic representations:

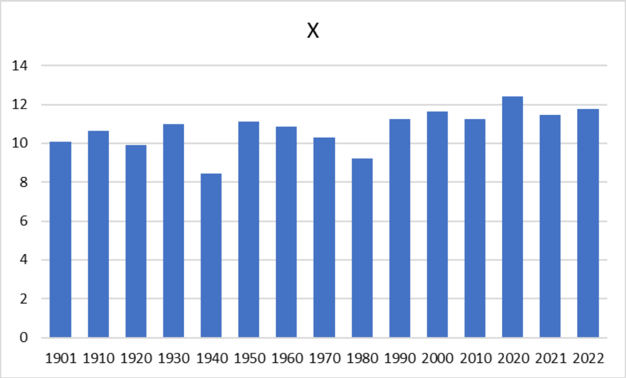
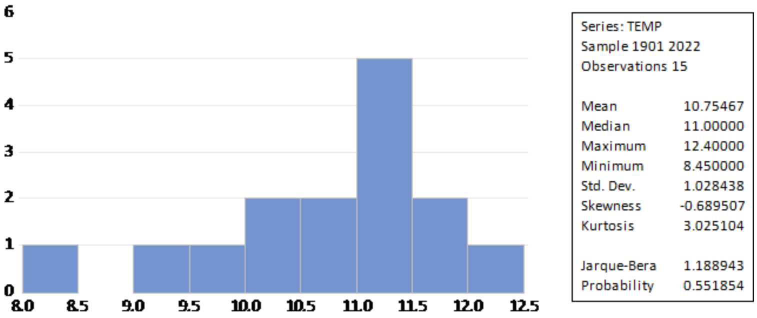


Figure 2. Evolution of the average annual air temperature in the period 1901-2022

As can be seen, both from the analysis of the series of data subjected to research, and especially from the figure presented above, in the considered time interval, the trend of the average annual temperature is increasing. In the main cities of Romania, temperatures have increased by at least 2°C in the last decades.

With the help of the Eviews 12 computer package, we performed a series of statistical tests aimed at providing a more accurate picture of the evolution of the average annual temperature in Romania during the considered period. Thus, we can note that the average value of this indicator for the time interval 1901-2022 is, with a variation between a minimum of $8,45°C$ (year 1940) and a maximum of $12,40°C$ (year 2020).



```
Series: TEMP
Sample 1901 2022
Observations 15

Mean         10.75467
Median       11.00000
Maximum      12.40000
Minimum      8.450000
Std. Dev.    1.028438
Skewness    -0.689507
Kurtosis     3.025104

Jarque-Bera  1.188943
Probability  0.551854
```

Figure 3. The main statistical tests performed on the average annual temperature value

The values of the previously performed statistical tests allow us to state that the distribution of the average annual temperature values for the considered interval is not perfectly symmetrical (the value of the skewness test is different from zero), the distribution being rather leptokurtic (kurtosis > 3). Moreover, it can be noted that, within the considered data series, the values between the minimum and the average of the series are less than those included in the second half of the variation range of the indicator subject to this research [5].

A similar analysis can be made for the evolution of annual precipitation over the time period 1901-2022. The main elements obtained from the analysis carried out using the Eviews 12 software can be presented as follows [7]:

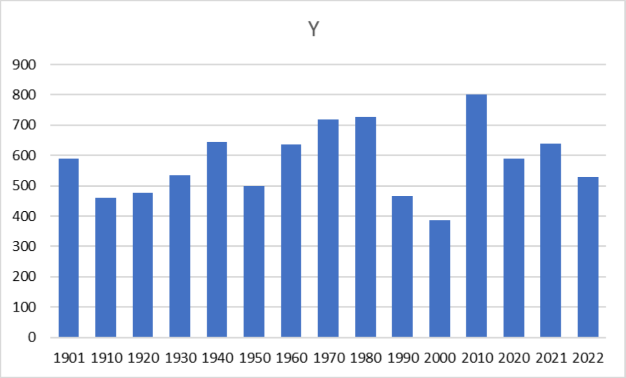
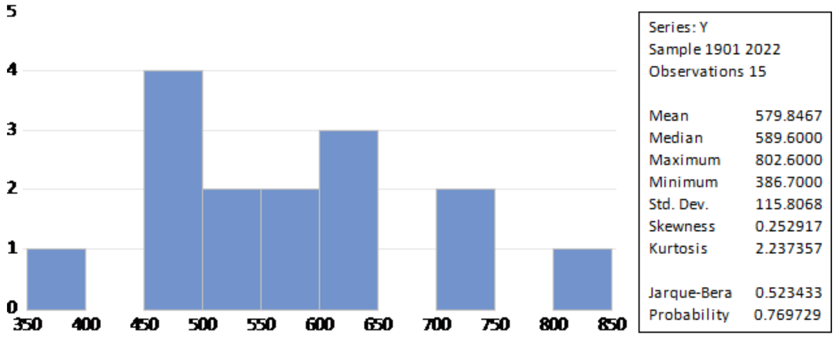


Figure 4. Evolution of the annual amount of precipitation in the time interval 1901-2022

The previous graphic representation allows us to affirm the fact that, during the time period subject to this research, the annual amount of precipitation had an oscillating trend in the period 1901-2022 with abundant precipitation in 2010. Similar to what was found in the case of the analysis of the average annual temperature, can observe the fact that the year 2000 is an anomaly compared to the general rule of evolution of the annual amount of precipitation in our country. Thus, it is noted that this is the only interval in which the value of this indicator decreases compared to the immediately preceding time period [3].



Series: Y
Sample 1901 2022
Observations 15

Mean	579.8467
Median	589.6000
Maximum	802.6000
Minimum	386.7000
Std. Dev.	115.8068
Skewness	0.252917
Kurtosis	2.237357
Jarque-Bera	0.523433
Probability	0.769729

Figure 5. Statistical tests performed on the annual amount of precipitation in the period 1901-2022

With the help of the Eviews program, we determined the range of variation of the researched indicator, establishing the fact that the value of the annual amount of precipitation falls between 386,7 mm, in the year 2000 and 802,6 mm, in the year 2010. We were also able to establish the fact that the average value of this indicator for the period 1901-2022 is 579,85 mm.

As can be seen, the values related to the Skewness and Kurtosis tests allow us to affirm the fact that the considered distribution is not a perfectly symmetrical one, predominating the values located between the minimum and the average of the data series, the distribution being rather platykurtic.

From the two analyzes carried out previously, it was possible to draw a very important conclusion regarding the method of analyzing the correlation between the two indicators subject to research, the average annual temperatures and the annual amounts of precipitation in Romania. Thus, it is noted that the evolution of the two climate indicators is not very similar. It can also be observed that the statistical tests performed on the data series related to the two indicators are not identical. Based on these findings, we can affirm the fact that there is a weak interdependence between the evolution of the average annual temperature and the annual amounts of precipitation.

To confirm this statement, as well as to identify the typology of the regression function, we created a graphic representation of the pairs of points that include the values of the average annual temperature and those of the corresponding annual amounts of precipitation. This graphic representation is as follows:

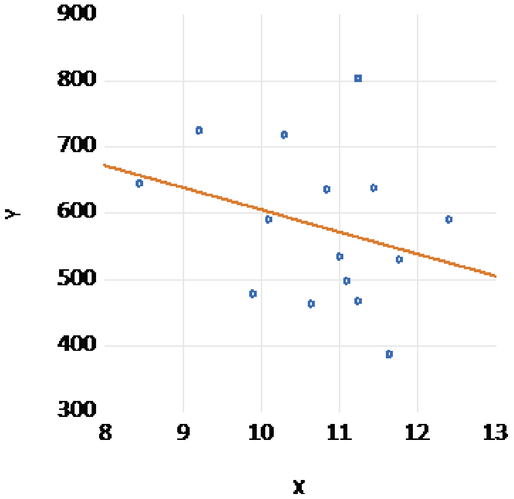


Figure 6. Correlation average annual temperature - annual amount of precipitation

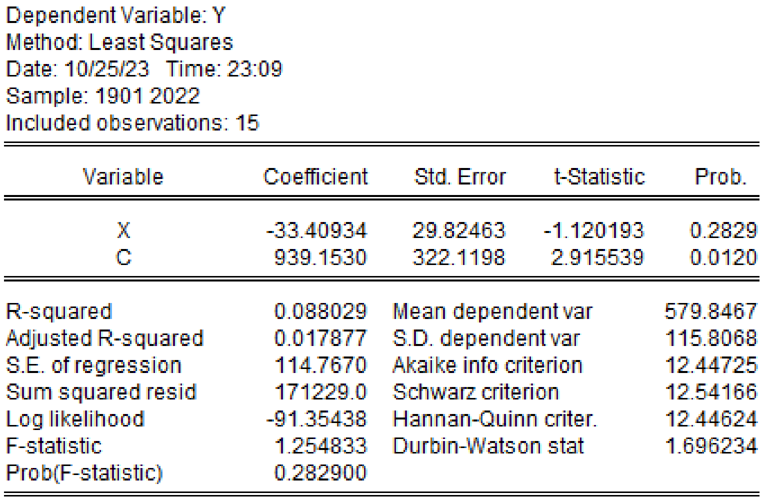
As can be seen from the previous graphic representation, the distribution of the empirical points of the average annual temperature and the annual amount of precipitation in Romania, in the period 1901-2022, suggests a linear and weak connection because the cloud of points is diffuse.

Analyzing the graph above, we find that the set of value pairs (x_i, y_i) reflects an inverse statistical relationship (as x_i is larger, y_i tends to be smaller).

The main problem of any regression model is the determination of the model parameters, an operation that can be performed with the help of the least squares method, as mentioned in paragraph 2.

In order to estimate the parameters of this regression model, we used the Eviews 12 computer program, within which we defined the equation that has as the result variable the value of the annual amount of precipitation, and as the factorial variable the value of the average annual temperature. We also considered the fact that this regression model will contain the free term c. The estimation method defined in the program is the least squares method.

Based on the previously presented elements, with the help of the Eviews 12 program, the following results were obtained [2]:



```
Dependent Variable: Y
Method: Least Squares
Date: 10/25/23   Time: 23:09
Sample: 1901 2022
Included observations: 15
==============================================================
        Variable      Coefficient   Std. Error   t-Statistic   Prob.
==============================================================
           X           -33.40934    29.82463    -1.120193    0.2829
           C            939.1530    322.1198     2.915539    0.0120
==============================================================
R-squared              0.088029   Mean dependent var     579.8467
Adjusted R-squared     0.017877   S.D. dependent var     115.8068
S.E. of regression     114.7670   Akaike info criterion  12.44725
Sum squared resid      171229.0   Schwarz criterion      12.54166
Log likelihood        -91.35438   Hannan-Quinn criter.   12.44624
F-statistic            1.254833   Durbin-Watson stat     1.696234
Prob(F-statistic)      0.282900
==============================================================
```

TABLE 2.The results of method Least Squares

The results of estimating the parameters of the regression model

Analyzing the previously obtained results, the following conclusions can be drawn:

- R-squared, the coefficient of determination, $R^2 = 0,088029$ shows that approximately 8,8% of the variation in the annual amount of precipitation can be explained by the level of the average annual temperature. It can be stated that the change in the average annual temperature is not a decisive factor in the variation of the annual amount of precipitation.

- For each independent and constant variable Eviews reports the standard error of the coefficient, the t-Statistic test and its associated probability. Working at the 5% level of relevance, as the probability attached to the *t-Statistic* test is the same for the annual average temperature variable, the n the coefficient is considered statistically significant.

The coefficient of the free term is not significant because the probability attached to the t-Statistic test is higher than the 5 % significance level.

In this context, we can extract from the results presented by the specialized computer program Eviews the following simple linear regression model:

$$y = -33,40934\ x + 939,1530$$

The negative sign of the regression coefficient confirms what was presented above, that is, the existence of an inverse relationship between the two studied variables.

5. Conclusions

One of the main chapters of statistics considers the possibility of making predictions. Although perfect relationships are not found in the real world, regression can be used to make forecasts of one variable, depending on the value of another.

Obviously, the results of this work are not deterministic, because the average annual air temperature and the annual amount of precipitation depend on many other factors, which cannot be taken into account, for example, the accelerated pollution of the atmosphere, water management, land use technologies, etc.

The global warming of the climate is considered not only the greatest meteorological risk, but also the greatest environmental risk, the negative consequences of which affect all the Earth's geospheres.

The large number of investigations in recent decades, both in the physical and biological spheres, and their interaction with regional and national climate change have made it possible to carry out a broader and firmer assessment of the interactions between the observed warming and its consequences.

The main conclusion is that many natural and artificial ecosystems are influenced by regional climate change, especially the increase in temperature and the intensification of natural risk phenomena.

The change in temperature and amount of precipitation will lead to changes in vegetation periods, the hydrological regime of rivers, soil erosion, floods, droughts and extremely strong torrential rains.

Namely, by making certain predictions, we can realize the seriousness of the situation in which we can find ourselves at a moment, in order to have the possibility to prevent some facts and events, which are sometimes catastrophic, if the necessary measures are not taken at the right time.

The period 1991-2020, considered the current climatic reference period according to the recommendations of the World Meteorological Organization, registers an increase of $0{,}5^o$ C in the multi-annual average annual air temperature in Romania, compared to the previous period 1981-2010.

The causes of the artificial increase in temperature are multiple: physical differences between urban and rural areas, including the absorption of sunlight, increased heat storage by artificial surfaces, obstruction of re-radiation by buildings, absence of plant transpiration, differences in air circulation, massive expansion of black surfaces (asphalt on streets and roofs) that increase the absorption of solar radiation during the day and re-radiation during the night, as well as other phenomena [1].

References

[1] Aja, S. U., Cranganu, C., *Will New York Be a "Baked Apple"? Using Data Sets to Explore Climate Change, Exploring the Earth System*, Kendall Hunt Publishing Co., 297-304, 2017.

[2] Anghelache, C-tin, *Analysis of the correlation between GDP and final consumption, Theoretical and Applied Economics*, Vol. XVIII, No. 9(562), 84-93, 2011.

[3] Bewick, V., Cheek, L. & Ball, J., Statistics review 7: *Correlation and regression*, Crit Care 7, 451, 2003. https://doi.org/10.1186/cc2401

[4] Bretscher, O., *Linear Algebra With Applications*, 3rd ed. Upper Saddle River NJ: Prentice Hall, 1995.

[5] Joanes, D. N., Gill, C. A., *Comparing measures of sample skewness and kurtosis, Journal of the Royal Statistical Society: Series D (The Statistician)*, Vol. 47, 1, 183-189, April 1998. https://doi.org/10.1111/1467-9884.00122

[6] Nufang Fang, Zhihua Shi, Fangxin Chen, Yixia Wang, *Least Squares Regression for Determining the Control Factors for Runoff and Suspended* Sediment Yield during Rainfall Events, Water, 7(7), 3925-3942, 2015. https://doi.org/10.3390/w7073925

[7] Popescu, A., *Analysis of the evolution and correlation between gross net salary and consumer price index*, Transylvanian Journal Of Mathematics And Mechanics, 10,(2), 113-120, 2018.

[8] Solomon, S., *Climate Change The Physical Science Basis,* IPCC, AGU Fall Meeting Abstracts, December 2007. Bibcode: 2007AGUFM.U43D..01S

[9] Ţurcanu, A., *Application of the least squares method to study the correlation between climatic factors in the Republic of Moldova*, Revista Acta et commentaryes, Technical University of Moldova, No. 2, 138-143, 2017. http://repository.utm.md/handle/5014/10544

[10] https://insse.ro/cms/ro/content/anuarul-statistic-al-rom%C3%A2niei

[11] http://www.mmediu.ro/articol/2022-a-fost-al-treilea-cel-mai-calduros-an-din-istoria-masuratorilor-meteorologice-din-romania/5909

University of Petroşani
Department of Mathematics and Computer Science
Str. Universităţii, 20, 332006, Petroşani, Romania
E-mail address: am.popescu@yahoo.com