Acta et Commentationes, Exact and Natural Sciences
Volume 1(15), 2023, Pages 67–77
CZU: 551.577.38+632.112(478)
ISSN: 2537-6284
E-ISSN: 2587-3644
DOI: 10.36120/2587-3644.v15i1.67-77

Regression analysis in the process of studying the correlation between climate factors of the Chisinau weather station

Anatolie Puțuntică and Vitalie Puțuntică

Abstract. The study presents linear and nonlinear mathematical models that analyze the regression in the process of correlation between climatic factors. The elaborated research is carried out for the period 1960-2019, analyzing the experimental data of the average annual temperature and the amount of annual precipitation. Various forms of regressions (linear, parabolic, cubic, etc.) and predictions for the correlation between the year and the average annual temperature, the year and the amount of annual precipitation, the average annual temperature and the amount of annual precipitation were obtained.
Keywords: regression, covariance, correlation coefficient.

Analiza de regresie în procesul de studiere a corelației dintre factorii climatici ai stației meteo Chișinău

Rezumat. Studiul prezintă modele matematice liniare și neliniare care analizează regresia în procesul de corelație între factorii climatici. Cercetarea elaborată este realizată pentru perioada 1960-2019, analizând datele experimentale privind temperatura medie anuală și cantitatea de precipitații anuale. Au fost obținute diverse forme de regresii (liniare, parabolice, cubice etc.) și prognoze pentru corelația dintre an și temperatura medie anuală, an și cantitatea de precipitații anuale, temperatura medie anuală și cantitatea de precipitații anuale.
Cuvinte-cheie: regresie, covarianță, coeficient de corelație.

1. Introduction

The forms of manifestation of the interdependence relations between the processes and the natural phenomena are extremely varied and most often difficult to notice. An essential problem to be solved in the analysis of the link between a dependency variable (denoted by y) and one or more independent variables (denoted by x_i) is the existence of the link between them. In practice, the following situations are encountered [1]:

a) the independent variable determines the modification of the dependent variable in case there is an univocal;
b) the two variables influence each other;

c) the variables evolve similarly independently, but influenced by another variable simultaneously;

d) the variables have a similar evolution without any connection between them.

Thus, for the systematic study of the relations between the two types of variables it is necessary to classify them according to certain criteria:

- by the nature of the interdependence relationship (functional and statistical links);
- by the number of factorial variables (single and multiple links);
- by the nature of the characteristics (association and correlation links);
- by the direction of the connection (direct and inverse connections);
- according to the shape of the function by which the connection is described (linear and nonlinear connections);
- after the connection time (synchronous and asynchronous connections).

2. Analytical method of measuring links

Analytical methods are those that allow the precise determination of both the relationship between two or more variables and its intensity. The analytical methods are:

- regression method;
- correlation method.

2.1. Regression method

This method is based on the use of mathematical functions to describe the shape of the connection between variables. The regression function has the following form:

$$y = f(x_1,\ x_2,\ \ldots,\ x_n) + \varepsilon,$$

where y - the dependent variable; $x_1,\ x_2,\ \ldots,\ x_n$ - independent variables; n - the number of influencing factors; ε - random variable or error that synthesizes the influence of unspecified factors.

In relation to the number of registered influencing factors are [2]:

- simple regression (unifactorial);
- multiple regression (multifactorial).

Only simple regression will be used in this research. It is based on function $y = f(x)+\varepsilon$ and studies the variation of a dependent variable y in relation to a single independent variable x, the other factors being considered neglected and with constant action.

The choice of the function is made with the help of the correlation graph. The most common simple correlation functions used are:

(1) $y = ax + b$ (linear);
(2) $y = ax^2 + bx + c$ (parabolic);
(3) $y = ax^3 + bx^2 + cx + d$(cubic);
(4) $y = ax^4 + bx^3 + cx^2 + dx + e$ (polynomial of degree IV);
(5) $y = ax^b$ (power);
(6) $y = ab^x$ (exponential);
(7) $y = a + b\ln x$ (logarithmic);
(8) $y = \frac{1}{a+bx}$ (hyperbole);
(9) $y = \frac{ax}{x+b}$ (Törniquist),

where a, b, c, d, e, f are parameters to be determined.

To determine the parameters a, b, c, d, e, f, the least squares method is usually used, according to which in order for the chosen regression function to be really significant we must have:

$$S = \sum_{i=1}^{n} \left(y_i - y_{x_i}\right)^2$$

to be minimal, where $i = \overline{1, n}$ - number of statistical units observed, y_i – the observed (empirical) values of the dependent variable, y_{x_i} – the theoretical values expressed by the regression equation. Determination of the values of each parameter (a, b, c etc.) is done by applying the conditions for obtaining the minimum value in the partial derivatives of the function S considered in the variables a, b, c etc.:

$$\frac{\partial S(a,\ b,\ c\ ...)}{\partial a} = 0,\quad \frac{\partial S(a,\ b,\ c,\ ...)}{\partial b} = 0,\quad \frac{\partial S(a,\ b,\ c,\ ...)}{\partial c} = 0, \ldots\ .$$

The formulas deduced for the mentioned correlation functions are brought in Tab. 1

Table 1. Functions formula correlation

Linear	$y = ax + b$	$\begin{cases} a\sum x^2 + b\sum x = \sum xy, \\ a\sum x + bn = \sum y; \end{cases}$
Parabolic	$y = ax^2 + bx + c$	$\begin{cases} a\sum x^4 + b\sum x^3 + c\sum x^2 = \sum x^2 y, \\ a\sum x^3 + b\sum x^2 + c\sum x = \sum xy, \\ a\sum x^2 + b\sum x + cn = \sum y; \end{cases}$
Cubic	$y = ax^3 + bx^2 + cx + d$	$\begin{cases} a\sum x^6 + b\sum x^5 + c\sum x^4 + d\sum x^3 = \sum x^3 y, \\ a\sum x^5 + b\sum x^4 + c\sum x^3 + d\sum x^2 = \sum x^2 y, \\ a\sum x^4 + b\sum x^3 + c\sum x^2 + d\sum x = \sum xy, \\ a\sum x^3 + b\sum x^2 + c\sum x + dn = \sum y; \end{cases}$

Polynomial of degree IV	$y = ax^4 + bx^3 + cx^2 + dx + e$	$\begin{cases} a\sum x^8 + b\sum x^7 + c\sum x^6 + d\sum x^5 + e\sum x^4 = \sum x^4 y, \\ a\sum x^7 + b\sum x^6 + c\sum x^5 + d\sum x^4 + e\sum x^3 = \sum x^3 y, \\ a\sum x^6 + b\sum x^5 + c\sum x^4 + d\sum x^3 + e\sum x^2 = \sum x^2 y, \\ a\sum x^5 + b\sum x^4 + c\sum x^3 + d\sum x^2 + e\sum x = \sum xy, \\ a\sum x^4 + b\sum x^3 + c\sum x^2 + d\sum x + en = \sum y; \end{cases}$
Power	$y = ax^b$	$\begin{cases} n\lg a + b\sum \lg x = \sum \lg y, \\ \lg a\sum \lg x + b\sum (\lg x)^2 = \sum \lg x \cdot \lg y; \end{cases}$
Exponential	$y = ab^x$	$\begin{cases} n\lg a + \lg b\sum x = \sum \lg y, \\ \lg a\sum x + \lg b\sum x^2 = \sum x \lg y; \end{cases}$
Logarithm	$y = a + b\ln x$	$\begin{cases} an + b\sum \ln x = \sum y, \\ a\sum \ln x + b\sum \ln^2 x = \sum y \ln x; \end{cases}$
Hyperbole	$y = \frac{1}{a+bx}$	$\begin{cases} an + b\sum x = \sum 1/y, \\ a\sum x + b\sum x^2 = \sum x/y; \end{cases}$
Törniquist	$y = \frac{ax}{x+b}$	$\begin{cases} \frac{n}{a} + \frac{b}{a}\sum \frac{1}{x} = \sum \frac{1}{y}, \\ \frac{1}{a}\sum \frac{1}{x} + \frac{b}{a}\sum \frac{1}{x^2} = \sum \frac{1}{xy}. \end{cases}$

2.2. Correlation method

The correlation method is used to measure the intensity of the link between the dependent variable y and the independent variable x. Depending on the nature of the link between the dependent variable y and the independent variable x, the correlation can be positive (in the case of the direct link) or negative (in the case of the reverse link). In this method the following indicators are used: covariance, correlation coefficient and correlation ratio [1,2].

Covariance captures the existence and direction of the link between the dependent variable y and an independent variable x. It is calculated as the simple arithmetic mean of the products of the deviations of the two correlated variables y and x from their arithmetic average $\bar{y}$ and $\bar{x}$ using the relation:

$$\operatorname{cov}(x, y) = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{n}.$$

The positive values of this indicator reflect a direct link, and the negative ones reflect an inverse link.

High values of the indicator show a strong link, while values close to zero signify the lack of links between the variables y and x. The correlation coefficient is determined according to one of the relations:

$$r = \frac{n\sum x_i y_i - \sum x_i \cdot \sum y_i}{\sqrt{\left(n\sum x_i^2 - (\sum x_i)^2\right) \cdot \left(n\sum y_i^2 - (\sum y_i)^2\right)}} \Leftrightarrow r = \frac{\sum (x_i - \bar{x})\,(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}.$$

The correlation coefficient can take values between -1 and 1. If $r \in [-1, 0)$ – reverse link, $r \in [0,\ 1)$ – direct link, $r = 0$ – the two variables do not correlate linearly.

The correlation ratio measures the intensity of the connection between the dependent variable y and the independent variable x in the case of nonlinear regression functions. The correlation ratio is determined by the relation:

$$F = \sqrt{1 - \frac{\sum \left(y_i - y_{i,\ curba}\right)^2}{\sum (y_i - \bar{y})^2}}.$$

Thus, the value of the correlation ratio is always positive and between 0 and 1. In the case of linear type connections the correlation ratio must be equal to the correlation coefficient.

Calculating this coefficient F for each of the variants of the functions considered we obtain a value close to 1. The most appropriate function is the one with the value F closest to 1.

3. Study of the correlation between some climatic elements at the Chisinau Meteorological Station

In this study, we will study the correlation between: the calendar year and the annual temperature, the calendar year and the annual amount of precipitation, the annual temperature and the annual amount of precipitation. According to the State Hydro meteorological Service of Chisinau station [3,4], in Tab. 2 we indicated the average annual temperature ($^{\circ}C$) and the amount of annual precipitation (mm) in the period 1960-2019, here by MT we denoted the medium temperature, and by AP - the amount of precipitation.

We will describe the process, for example, for the function of the parabolic form between the average annual temperature and the amount of annual precipitation. For this model, we complete the following table:

According to Tab. 1 (parabolic model), we obtain:

Table 2. Average annual temperature and amount of annual rainfall in the period 1960-2019, Chisinau

Year	MT	AP	Year	MT	AP	Year	MT	AP	Year	MT	AP
1960	10,575	537	1975	10,808	484	1990	11,342	361	2005	10,492	638
1961	10,408	450	1976	8,342	600	1991	9,433	673	2006	10,208	564
1962	10,092	559	1977	9,525	464	1992	10,117	417	2007	12,042	480
1963	9,192	532	1978	8,725	563	1993	9,408	532	2008	11,308	466
1964	9,442	511	1979	9,783	684	1994	11,342	415	2009	11,408	446
1965	9,033	537	1980	8,300	712	1995	10,017	702	2010	10,558	734
1966	10,858	774	1981	9,667	536	1996	9,050	711	2011	10,450	428
1967	10,042	481	1982	9,783	384	1997	9,400	607	2012	11,217	522
1968	9,992	532	1983	10,458	549	1998	10,250	668	2013	11,083	531
1969	8,692	525	1984	9,167	669	1999	11,025	485	2014	10,917	604
1970	10,067	672	1985	8,000	591	2000	11,150	437	2015	11,983	431
1971	9,967	590	1986	9,625	402	2001	10,308	618	2016	11,225	644
1972	9,758	621	1987	8,075	592	2002	10,842	604	2017	11,208	635
1973	9,500	396	1988	9,025	652	2003	9,800	459	2018	11,200	609
1974	9,958	562	1989	10,933	460	2004	10,300	591	2019	12,225	403

Table 3. Parabolic function model

n	x	y	x^2	x^3	x^4	xy	x^2y
1	10,575	537	111,830625	1182,608859	12506,08869	5678,775	60053,04563
2	10,408	450	108,333333	1127,569084	11736,11112	4683,7485	48750,00003
...	...	...	...	...	...	...	...
59	11,2	609	125,44	1404,928	15735,1936	6820,8	76392,96
60	12,225	403	149,450625	1827,033891	22335,48931	4926,675	60228,60188
Σ	609,1	33036	6240,452323	64502,06588	672353,4868	333539,6911	3398332,086

$$\begin{cases} 672353,486759a + 64502,065881b + 6240,452323c = 3398332,085749; \\ 64502,065881a + 6240,452323b + 609,1c = 333539,69105; \\ 6240,452323a + 609,1b + 60c = 6240,452323; \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} a = & -8,8035497; \\ b = & 145,4784458; \\ c = & -10,6131547. \end{cases}$$

Therefore, the parabolic regression will be:

$$y = -8,8035497x^2 + 145,4784458x - 10,6131547.$$

If the connection between the average annual temperature and the amount of average annual precipitation is assumed to be linear, then the coefficients of the equation $y = ax+b$ based on the system of equations must be determined using the least squares method (see Tab.1):

$$\begin{cases} 6240,4523236a + 609,1b = 333539,69105; \\ 609,1a + 60b = 33036; \end{cases} \Longleftrightarrow \begin{cases} a = -32,0781457; \\ b = 876,2466428. \end{cases}$$

So the regression line will be $y = 876,2466428 - 32,0781457x$. For the covariance sizes and the correlation coefficient of the regression line we will have to complete the table:

Table 4. Calculating covariance and regression.

n	x	y	$x-\overline{x}$	$(x-\overline{x})^2$	$y-\overline{y}$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
1	10,575	537	0,423333	0,179211	-13,6	184,96	-5,757333
2	10,408	450	0,256663	0,065876	-100,6	10120,36	-25,820331
...	...	...	...	...	...	...	...
59	11,2	609	1,048333	1,099003	58,4	3410,56	61,222667
60	12,225	403	2,073333	4,298711	-147,6	21785,76	-306,024
Σ	609,1	33036	0	57,07216	0	584508,4	-1830,76895

As a result we get:

$$r = \frac{\Sigma(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\Sigma(x_i - \overline{x})^2 \cdot \Sigma(y_i - \overline{y})^2}} = \frac{-1830,76895}{\sqrt{57,072156 \cdot 584508,4}} = -0,3169757358483542.$$

Because $\mathrm{cov}(x, y) < 0$, respectively $r < 0$, then we have a strong inverse link between the average annual temperature and the amount of average annual rainfall, well increasing the average annual temperature leads to a decrease in the amount of average annual rainfall.

To determine the correlation ratio for parabolic regression:

$$y_{parabola} = -8,8035497x^2 + 145,4784458x - 10,6131547,$$

we complete Table 5.

Table 5. Correlation for the parabolic regression.

x_i	y_i	$y_{i,parabola}$	$y_i - y_{i,parabola}$	$(y_i - y_{i,parabola})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
10,575	537	543,3149	-6,31495	39,87855	-13,6	184,96
10,408	450	549,8566	-99,8566	9971,347	-100,6	10120,36
...	...	...	...	...	...	...
11,2	609	514,4282	94,57183	8943,832	58,4	3410,56
12,225	403	452,1648	-49,1648	2417,182	-147,6	21785,76
Σ	33036	33036	-25,6656	519295,7	0	584508,4

So,

$$F = \sqrt{1 - \frac{519295,65987597423}{584508,4}} = 0,3340187496937681.$$

The graphical representation of the initial data, the parabolic regression and the linear regression is represented in Fig. 1.

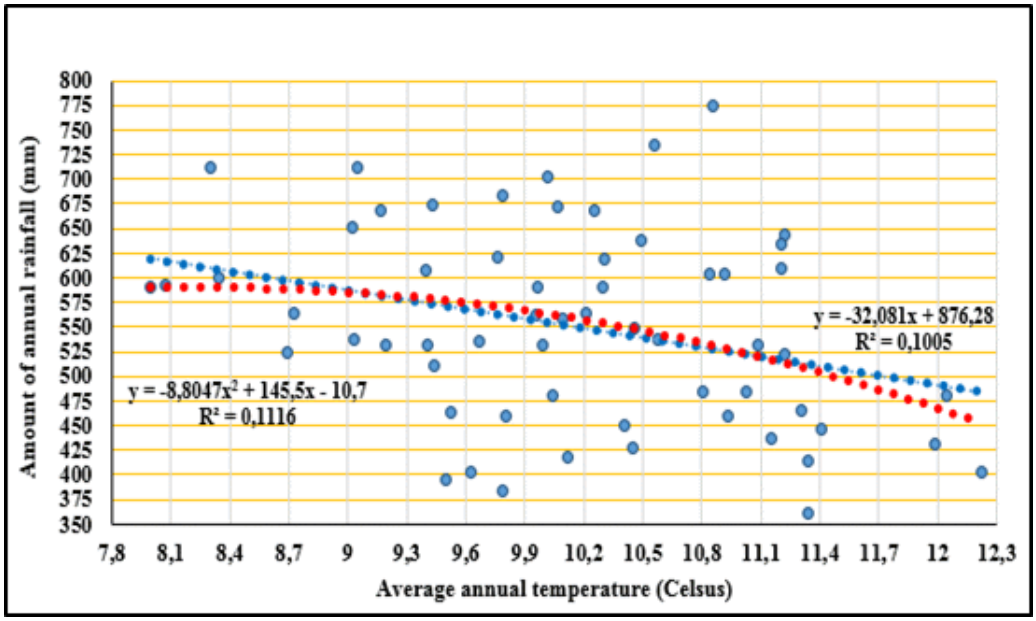

Figure 1. Linear regression (blue) and parabolic regression (red).

Correlation	Regression curve	Regression of curve coefficients	Correlation ratio	Forecast for 2020
Year and average annual temperature (°C)	linear	0,031261439	0,555102	11,1051
		-52,04296569		
	parabolic	0,001411821	0,677351	11,9951
		-5,586376217		
		5535,678664		
	cubic	$-2{,}8101 \cdot 10^{-5}$	0,687447	11,6603
		0,169132863		
		-339,2522238		
		226791,6331		
	quartic	$-2{,}56495 \cdot 10^{-7}$	0,687637	11,6044
		0,00201309		
		-5,922092462		
		7739,21872		
		-3790846,75		
	exponential	0,018015157	0,564372	11,1691
		1,003188088		
	logarithmic	-460,9379324	0,553585	11,0976
		62,02137463		
	hyperbole	0,741389957	0,573484	11,2382
		-0,000322974		
	Törniquist	$6{,}45085 \cdot 10^{14}$	0,303971	10,3113
		$1{,}26373 \cdot 10^{17}$		
Year and amount of annual rainfall (mm)	linear	-0,122422895	0,02148	546,866
		794,1603501		
	parabolic	-0,01772693	0,052732	535,692
		70,41303189		
		-69365,6667		
	cubic	0,000725385	0,06064	544,334
		-4,347184658		
		8683,477728		
		-5780743,873		
	quartic	-0,000120363	0,096406	518,114
		0,958571053		

		-2862,705381		
		3799573,226		
		-1891091005		
	exponential	854,0270487	0,021398	546,903
		0,999779385		
	logarithmic	2384,28375	0,021292	546,918
		-241,4139201		
	hyperbole	0,001025088	0,021316	546,94
		$3{,}97657 \cdot 10^{-7}$		
	Törniquist	385,6266587	0,02094	547,042
		-596,0389292		
Average annual temperature and amount of annual rainfall (mm)	linear	-32,07814573	0,316976	520,016
		876,2466428		
	parabolic	-8,803549677	0,334019	467,739
		145,4784458		
		-10,61315474		
	cubic	-11,75781748	0,374376	486,022
		346,625388		
		-3407,48275		
		11730,63422		
	quartic	-0,907613644	0,374624	494,343
		25,02808731		
		-208,9526932		
		297,6295967		
		2526,383239		
	exponential	977,2843906	0,313286	518,977
		-0,056666872		
	logarithmic	1276,94661	0,310794	521,142
		-314,038121		
	hyperbole	0,000806208	0,309868	518,131
		0,0000999988		
	Törniquist	365,1741364	0,297274	542,413
		-3,369321499		

Table 6. Correlations results for other types of regression.

From Fig.1 we notice that the increase of temperature leads to a decrease in the amount of precipitation. Therefore, we can make the following predictions: for an average annual air temperature of $11,015^\circ$ C, we will have according to the parabolic regression the amount of annual precipitation $519,257$ *mm* and according to the linear regression $520,014$ *mm*.

Proceeding according to the above model, the other types of regressions are also studied: cubic, degree IV curve, logarithmic, exponential, etc. The results are presented in Tab. 6.

Also, the correlation between the year and the average annual temperature, the year and the amount of annual precipitation are analyzed (see Tab. 6).

To estimate the amount of average annual rainfall for 2020, the average annual temperatures were obtained by correlating the year with the average annual temperatures.

4. Conclusions

As a result of analyzing all forms of correlation (calendar year and annual temperature, calendar year and annual rainfall, annual temperature and annual rainfall) it can be concluded that the best function is grade IV. Cubic function as well provides good results. In addition to the correlations of mathematical functions, obviously, we will take into account the geographical conditions of the weather station location (altitude, latitude, exposure, vegetation, transparency of the atmosphere, pollution, fragmentation of the relief, etc.). Similar works in the future could take into account other logical correlations, passed through the mathematical apparatus (temperature and humidity, cloudiness and precipitation, cloudiness and visibility, visibility and humidity - very important correlation for road, air, sea, river, rail transport traffic).

References

[1] Ciumac, P., Ciumac, V., Ciumac, M. *Teoria probabilității și elemente de statistică matematică*. Chișinău, Ed. Tehnică, UTM, 2003;

[2] Rancu, N., Tövissi, L. *Statistica matematică cu aplicații în producție*, Ed. Academiei Republicii Populare Române, 1963;

[3] www.meteo.md;

[4] www.bns.md.

Received: February 20, 2023 *Accepted: July 27, 2023*

(Anatolie Puțuntică, Vitalie Puțuntică) "Ion Creangă" State Pedagogical University of Chisinau
E-mail address: aputuntica@gmail.com, putunticavitalie@gmail.com