The method of correlation analysis in agriculture

V. V. Vakhabov*, and *M. A.* Hidoyatova

"Tashkent Institute of Irrigation and Agricultural Mechanization Engineers" National Research University, Tashkent, Uzbekistan

Abstract. The mass data obtained from scientific and practical experiments are mostly probable-random. The methods of mathematical statistics are used for their processing; they include correlation, regression methods, dispersion analysis methods, and others. In this paper, we propose a correlation-regression analysis of the results of an experiment using a specific example from agriculture. The processing method and analysis of experimental results described in the paper are scientific-methodological and will be useful for the specialists involved in scientific research.

1 Introduction

Mathematical analysis of the results of scientific research and the derivation of appropriate theoretical and practical conclusions are the most important issues for each experimenter (doctoral and graduate students, students - holders of master's degrees). To carry out these studies, one needs to know how to analyze the obtained experimental data. In many cases, the experimenter in his studies should be able to determine and evaluate the dependence of the calculated value on one or several random variables.

The relationship between features can be functional (complete) and correlational (statistical).

A functional relationship is a relationship between features in which each value of one variable (argument) corresponds to a strictly defined value of another variable (function).

Such connections are observed in mathematics, chemistry, physics, astronomy, and other sciences. For example, the area of a circle ($S = \pi R^2$) and the circumference ($C = 2\pi R$) are completely determined by the radius, the area of a triangle by its sides, and so on.

In socio-economic phenomena, functional relationships between attributes are rare; here, such relationships between variables occur more often, in which several values of others correspond to the numerical value of one of them. Such a relationship between traits is called a correlation (statistical) relationship; for example, it is known that the yield depends on the amount of fertilizer applied, but other factors also influence it (soil quality, precipitation, etc.). In addition, the same doses of fertilizers, ceteris paribus, often affect yields differently.

The correlation relationship is incomplete if it appears with many observations when

* Corresponding author: v.vaxobov2019@gmail.com

comparing the average values of the effective and factor signs; the corresponding mathematical equations express it.

There are rectilinear and curvilinear, direct and inverse, simple (measuring relationships between two features), and multiple (measuring relationships between three or more features) correlations. Using the correlation analysis method, two main tasks are solved: determining the forms of the constraint equation's parameters and measuring the connection's tightness. The first task is solved by finding the connection equation and determining its parameters, and the second - using various indicators of the tightness of the connection (correlation coefficient, correlation index, etc.)

Schematically, correlation analysis can be divided into five stages:

1) Statement of the problem, establishing the presence of a connection between the studied features;

2) Selection of the most significant factors for analysis;

3) Determining the nature of the relationship, its direction and form, the selection of a mathematical equation to express significant relationships;

4) Calculation of the numerical characteristics of the correlation relationship (determination of the parameters of the equation and indicators of the tightness of the relationship)

5) Statistical evaluation of selective communication indicators.

Choosing one or another equation for studying the relationships between features is the most difficult and crucial moment in correlation analysis. The mathematical relationship equation can be established with paired correlation by plotting (correlation field, etc.), compiling correlation tables, and revising various functions. In economic research, a straight-line form of relationship is often considered, which is expressed by a straight-line equation $y_x = a + bx$, where: y_x is equalized values of the resulting attribute (dependent variable) ; x is value of the factor sign (independent variable); a is the starting point, or the value y_x at $x = 0$ (it makes no economic sense) b is the regression coefficient, always a named number. If $b > 0$, the connection is direct, if $b < 0$, then the connection is inverse, if $b = 0$, there is no connection.

An equation of this type is called a regression equation or correlation dependence; its main task is to establish a quantitative relationship between features [3].

Equation parameters a and b are determined by the least squares method, which makes it possible to find such a theoretical regression line, which, compared with others, passes closest to the points of the correlation field representing the actual data, i.e., gives the smallest sum of squared deviations of the actual values of the resulting feature from the leveled (theoretical) values: $\sum(y_i - \bar{y}_k)^x = min.$

The procedure for obtaining a system of normal equations for pairwise correlation is as follows. To obtain the first equation of the system, it is necessary to multiply all the terms of the original correlation equation by the coefficient at the first unknown (a) and sum the resulting products. Then, to obtain the second equation, it is necessary to multiply all the terms of the original equation by the coefficient in the second unknown (b) and sum all the products. The technique for obtaining a system of normal equations remains the same for constructing a system of equations with a large member of variables. So, for a paired linear connection, the system of normal equations has the form:

$$\begin{cases} \sum y = an + b\sum x, \\ \sum yx = a\sum x + b\sum x^2 \end{cases} \quad (1)$$

The parameters a and b of the straight-line equation can be determined by other working formulas:

$$b = \frac{n\sum xy - \sum y \sum x}{n\sum x^2 - \sum x \sum x}; a = \frac{\sum y \sum x^2 - \sum yx \sum x}{n\sum x^2 - \sum x \sum x} \quad (2)$$

or

$$b = \frac{n\overline{xy} - \overline{x}\overline{y}}{n\bar{x}^2 - \bar{x}^2}; a = \bar{y} - b\bar{x}.$$

With a curvilinear dependence, the system of equations is constructed like for a linear dependence. Thus, the system of equations of the parabola $\bar{y}_x = a + bx + cx^2$ has the form.

$$\begin{cases} \sum y = an + b\sum x + c\sum x^2, \\ \sum yx = a\sum x + b\sum x^2 + c\sum x^3, \\ \sum yx^2 = a\sum x^2 + b\sum x^3 + c\sum x^4, \end{cases} \quad (3)$$

Correlation equations are used to calculate the theoretical regression line and the expected values of the dependent variable at the respective values of the factor(s). When studying the correlation, it becomes necessary, along with the solution of the regression equation, to also measure the degree of closeness of the relationship between the signs, which is characterized by a special relative indicator called the correlation coefficient.

With a paired linear dependence, the tightness of the connection is determined using the linear correlation coefficient:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y},$$

$$\overline{xy} = \frac{\sum xy}{n}\ ;\ \bar{x} = \frac{\sum x}{n};\ \bar{y} = \frac{\sum y}{n};\ \sigma_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2};\ \sigma_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2} \quad (4)$$

The linear correlation coefficient is used to assess the degree of closeness of the relationship with a linear relationship. For cases of a non-linear relationship between features, another formula for the correlation coefficient is used, which follows from the rule for adding variances:

$$\sigma^2_{general} = \sigma^2_{factor} + \sigma^2_{residual} \quad (5)$$

It can be seen from the above equality that the greater the influence of the factor on the resultant attribute, the more the value of its variance (σ^2_{factor}) approaches the value of the total variance of the resultant attribute. Accordingly, the more (σ^2_{factor}) and $\sigma^2_{residual}$ less, the closer the relationship between the features will be and vice versa. Because of this, the ratio of factorial and total variances are used to assess the closeness of the relationship between features. The correlation coefficient is calculated by the formula in the form:

$$r = \sqrt{\frac{\sigma^2_{factor}}{\sigma^2_{general}}} \quad (6)$$

Combination that $\sigma^2_{general} = \sigma^2_{factor} + \sigma^2_{residual}$, the correlation coefficient formula can be represented in another form:

$$r = \sqrt{1 - \frac{\sigma^2_{residual}}{\sigma^2_{general}}} \quad (7)$$

Both correlation coefficient formulas are applicable for calculating the tightness of the

connection for any form of connection. The correlation coefficient is in the range from 0 to± 1, if the correlation coefficient is equal to zero, then there is no connection, and if it is equal to one, then the connection is functional; the sign $\pm$ at the correlation coefficient indicates the direction of the connection (" + " - direct, " − " - reverse). The closer the correlation coefficient is to one, the closer the relationship between the features. The square of the correlation coefficient is called the coefficient of determination(r^2). It shows what proportion of the total variation of the trait is determined by the studied factor.

Now let's formulate tasks for applying the method of correlation analysis. This paper is devoted to determining the connection indices of curvilinear dependence for the problem in agriculture, described below.

The harvest yields of winter wheat from seven farms of the area were compared based on the prime cost of 1 centner of the grain of this crop. The results are shown in Table 1.

Table 1. Relation between the yield and prime cost

X– the yield (centner/ha)	8	11	13	19	21	27	29
Y– the prime cost for 1 centner (roubles.)	12	8	7.3	6.0	6.3	5.8	5.2

2 Research method

To study this problem, correlation analysis and the least squares method were used to establish the form, parameters of the equation of connection, and the tightness of connection between the random variables under consideration.

3 Results of the study

1. To determine the dependence of the prime cost (*y*) on productivity (*x*), we plot a graph of the correlation field (picture 1): the value of the factor attribute of an independent variable (yield) is set on the abscissa axis, and the resultant characteristic (dependent variable –the prime cost) - on the ordinate axis [3, 4].

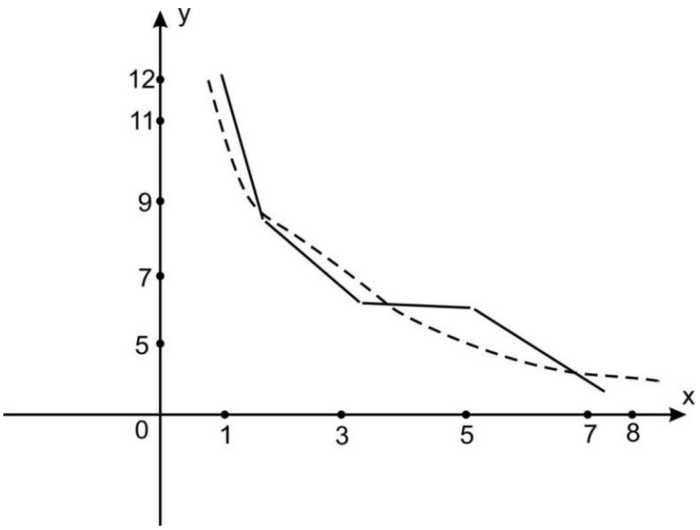

Fig. 1. The graph shows that, in this case, the connection is close to hyperbole, and the second-order

hyperbole equation can express it

$$y = a + \frac{b}{x^2} \qquad (8)$$

The solution of this regression equation shows the change in the prime cost under the influence of the yield eliminating the random fluctuations of the attribute.

2. To determine the parameters *a* and *b* of this equation, the following system of normal equations is used:

$$\begin{cases} an + \text{в} \sum \frac{1}{x^2} = \sum y \\ a \sum \frac{1}{x^2} + b \sum \frac{1}{x^4} = \sum \frac{y}{x^2} \end{cases}$$

The solution of this system with respect to parameters *a* and *b* leads to the following formulas:

$$a = \frac{1}{D}\left(\sum y \sum \frac{1}{x^4} - \sum \frac{y}{x^2} \sum \frac{1}{x^2}\right)$$
$$b = \frac{1}{D}\left(n \sum \frac{y}{x^2} - \sum y \sum \frac{1}{x^2}\right)$$

Where, $D = n \sum \frac{1}{x^2} - \left(\sum \frac{1}{x^2}\right)^2$

To determine the parameters *a* and *b*, we should first calculate $\sum y$, $\sum \frac{y}{x^2}$, $\sum \frac{1}{x^2}$and$\sum \frac{1}{x^4}$. For this, the following calculation Table 2 is drawn up.

Table 2. Calculaitoni summary table for different X values

X	x=$\frac{x}{8}$	y	$\frac{y}{x^2}$	$\frac{1}{x^2}$	$\frac{1}{x^4}$	$\overline{y_x}$
8	1.0	12.0	12.00	1.0000	1.0000	11.7
11	1.4	8.0	4.08	0.5102	0.2603	8.4
13	1.6	7.3	2.85	0.3906	0.1526	7.6
19	2.4	6.0	1.04	0.1736	0.0301	6.1
21	2.6	6.3	0.93	0.1479	0.0219	5.9
27	3.4	5.8	0.50	0.0865	0.0075	5.5
29	3.6	5.2	0.40	0.0772	0.0060	5.4
Total:	-	50.6	21.80	2.3860	1.4784	50.6

In this table, *X* indicates the yield of wheat (centner/ha) in different farms of the area, and *Y* indicates the prime cost of 1 centner of wheat (roubles). To simplify the calculation of auxiliary quantities, the value of independent variable *X* is reduced by 8; the results are placed in the second column (*x*) of Table 2. Using the sums from Table 2, we find the values of the system determinants:

$$D = n \sum \frac{1}{x^4} - \left(\sum \frac{1}{x^2}\right)^2 = 7 \cdot 1.4784 - (2.386)^2 = 10.3488 - 5.9630 = 4.6558;$$

$$A = \sum y \sum \frac{1}{x^4} - \sum \frac{y}{x^2} \sum \frac{1}{x^2} = 50.6 \cdot 1.4784 - 21.80 \cdot 2.3860 = 22.7922;$$

$$B = n\sum\frac{y}{x^2} - \sum y\sum\frac{1}{x^2} = 7 \cdot 21.80 - 50.6 \cdot 2.3860 = 318684.$$

Hence

$$a = \frac{A}{D} = \frac{22.77}{4.65} = 4.895 \qquad b = \frac{B}{D} = \frac{31.868}{4.656} = 6.8445$$

Thus, the empirical equation of the second-order hyperbola is as follows:

$$\overline{y_x} = 4.9 + \frac{6.8}{x^2}$$

The expected values of the dependent variable $\overline{y_x}$ calculated by this equation are shown in the last column of Table 2.

It can be seen that they are in good agreement with the empirical values of the У attribute. This can also be seen from the figure, which shows the lines of regression of empirical and flattened ones by the second-order hyperbole equation.

However, in practice, there are cases when, with an increase in the independent variable X, the dependent variable У, rapidly decreasing, soon stabilizes at a certain level, taking more or less constant values. In such cases, to flatten the empirical regression series, you can use the third-order hyperbola equation in the form [7], [8]:

$$\bar{y}_x = a + \frac{b}{x^3} \qquad (9)$$

To determine the parameters$a$ and$b$ of this equation, the following system of normal equations is used:

$$\begin{cases} an + b\sum\frac{1}{x^3} = \sum y \\ a\sum\frac{1}{x^3} + b\sum\frac{1}{x^6} = \sum\frac{y}{x} \end{cases}$$

Solving these equations together for the parameters, a and b we have the following formulas

$$a = \frac{1}{D_1}\left(\sum y\sum\frac{1}{x^6} - \sum\frac{y}{x^3}\cdot\sum\frac{1}{x^3}\right);$$

$$b = \frac{1}{D_1}\left(n\sum\frac{y}{x^3} - \sum y\cdot\sum\frac{1}{x^3}\right)$$

Where $D_1 = n\sum\frac{1}{x^3} - \left(\sum\frac{1}{x^3}\right)^2$ it is obvious that to find the parameters a and b, it is necessary to first calculate$\sum y$, $\sum\frac{y}{x^3}\sum\frac{1}{x^6}$.

The following Table 3 shows the results of the eight tests of the same types and their processing according to the formula (2).

Table 3. Calculaiton results of 8 tests according to formula 2

x	y	x^3	x^6	$\frac{y}{x^3}$	$\frac{1}{x^3}$	$\frac{1}{x^4}$	$\bar{y}_x$
1	29.0	1	1	29.0	1.0000	1.00000	28.4
2	5.9	8	64	0.738	0.1250	0.01562	5.7
3	3.4	27	729	0.126	0.0370	0.00137	3.4
4	3.8	64	4096	0.059	0.0156	0.00024	2.9
5	2.5	125	15625	0.020	0.0080	0.00006	2.7
6	2.0	216	46656	0.009	0.0046	0.00002	2.6
7	2.3	343	117649	0.007	0.0029	0.00001	2.6
8	1.9	512	262144	0.004	0.0020	0.00000	2.5
sum	50.8	-	-	29.963	1.1951	1.01732	50.8

It can be seen from the data in this table that after a sharp decrease in the numerical values of У, they gradually stabilize, remaining about at the same level. Let's find the empirical equation for this regression:

1) $D = 8 \cdot 1.01732 - (1.195)^2 = 8.1386 - 1.4283 = 6.710;$
2) $a = \frac{1}{6.710}[50.8 \cdot 1.01732 - 29.963 \cdot 1.1951] = \frac{15.871}{6.710} = 2.4;$
3) $b = \frac{1}{6.710}[8 \cdot 29.963 - 50.8 \cdot 1.1951] = \frac{178.993}{6.710} = 26.4.$

And so, the empirical regression equation У for X has the form

$$\bar{y}_x = 2.4 + \frac{26.4}{x^3}$$

The equalizing values of $\bar{y}_x$ calculated by this equation are given in the last column of Table 3. It can be seen that they agree with the empirical values of the variable, y.

Sometimes, when studying the relationship between the values of *X* and *Y*, it will be necessary to apply a regression expressed by the first-order hyperbola equation[9-14], with three unknowns, for example, *a, b,* and *c*. If, with an increase in the independent variable X, the dependent variable *Y* decreases rapidly, reaching a certain limit beyond which a more or less stable flow of the function is found, then the hyperbola equation of the following form can be used to level the empirical values of the dependent variable:

$$y = f + bx + \frac{c}{x} \qquad (3)$$

To determine the parameters a, b, c of this equation, the following system of normal equations is used:

$$\begin{cases} an + b\sum x + c\sum \frac{1}{x} = \sum y; \\ a\sum x + b\sum x^2 + nc = \sum xy \\ a\sum \frac{1}{x} + bn + c\sum \frac{1}{x^2} = \sum \frac{y}{x} \end{cases} \qquad (3)$$

To compose a system using sample data, it is necessary to first calculate $\sum x$, $\sum y$, $\sum xy$, $\sum \frac{1}{x}$, $\sum \frac{1}{x^2}$.

Example - № 3. As shown by numerous observations, with an increase in the number of independent trials (n), the error value of the average result $S\bar{x}$ naturally decreases

Table 4. Average errors of different numbers of tests

Number of tests *n*	5	10	15	20	25	30	35	40	45
Average error $S\bar{x}$	6.2	2.9	1.6	1.9	1.1	0.9	1.2	0.9	0.9

This connection between the variables *n* and $S\bar{x}$ has a hyperbolic character and can be described using equation (3). To simplify the computational work, we denote the variables n and$S\bar{x}$, respectively, by *X*and *Y,* we will reduce the values of the independent variable *X* by *S*. Then the argument *X* will be expressed as a series of natural numbers 1, 2, 3,…,9. Now let's calculate Table 4.

Table 5. Summery of calculations for X values

X	У	XУ	$\frac{У}{X}$	X^2	$\frac{1}{X}$	$\frac{1}{X^2}$	$\bar{У}_x$
1	6.2	6.2	6.200	1	1.000	1.0000	6.2
2	2.9	5.8	1.450	4	0.500	0.2500	2.9
3	1.6	4.8	0.533	9	0.333	0.1111	1.9
4	1.9	7.5	0.475	16	0.250	0.0625	1.4
5	1.1	5.5	0.220	25	0.200	0.0400	1.2
6	0.9	5.4	0.150	36	0.167	0.0278	1.1
7	1.2	8.4	0.171	49	0.143	0.0204	1.0
8	0.9	7.2	0.113	64	0.125	0.0156	1.0
9	0.9	8.1	0.100	81	0.111	0.0123	0.9
$\sum$ 45	17.6	59.0	9.412	285	2.829	1.5397	17.6

Based on the results of this table 5, we compose a system of normal equations (4):

$$\begin{cases} 9a + 45b + 2.83c = 17.6 \\ 45a + 28b + 9c = 59.0 \\ 2.83a + 9b + 1.54c = 9.412 \end{cases}$$

Solving this system concerning the parametersa, b, c, we find:

$$a = -0.571; b = 0.0871; c = 6.6496$$

Then the empirical regression equation for *Y by* X, has the form:

$$\bar{У}_x = -0.571 + 0.0871 + \frac{6.6496}{x}$$

Comparing У with $\bar{У}_x$, we note that it agrees well with the empirical values of the function. In conclusion, we note that in other similar cases, for leveling regression series, the hyperbola equations of the second and third orders, with three unknowns, may be more suitable.

$\bar{Y}_x = a + bx + \frac{c}{x^2}$; $\bar{Y}_x = a + bx + \frac{c}{x^3}$; and then they gave.

4 Conclusions

1) The regression equations characterizing the yield and prime cost connection are derived.

2) It is determined that as the yield increases, the prime cost stabilizes around the parameter value of a=4.9.

3) In the case when the results of the experiment show that with the increase in X, the dependent variable У decreases rapidly, then it is convenient to use the third-order hyperbola equation to flatten the empirical series.

References

1. Aderhold J, Davydov V Yu, Fedler F, Klausing H, Mistele D, Rotter T, Semchinova O, Stemmer J and Graul J 2001 J. Cryst. Growth **222** 701
2. Gmurman V Ye 1977Probability theory and mathematical statistics (Moscow: High school) p 480
3. Lozinskiy S N1975 Collection of problems on probability theory and mathematical statistics(Moscow: Statistics) p 198
4. Baraz V.R. 2005 Correlation-regression analysis of the relationship between indicators of commercial activity using the Excel program.(Yekaterinburg) pp11-25
5. Gataulin A M, Kharitonova L and Gavrilova GV 1986 Economic and mathematical methods in agricultural production planning(Moscow: Kolos) p 330
6. Boyarskiy A L1957 Mathematics for economists (Moscow) p 267
7. Vakhobov V and Khidoyatova M2018 J.Irrigation and melioration**4 (14)**
8. Vakhobov V and Khidoyatova M "On the processing of experimental date by methods of mathematical statistics" EPRA International of Agriculture and Rural Economic Research(ERER) Peer-Reviewed Journal Volume:8 |Issue:3| September 2020
9. Fedoseev VV2000 Economics - mathematical methods and applied models (Moscow) p 345
10. Kremer N M2004 Higher Mathematics for Economists(Moscow) p 340
11. Abdullayev A, Kholturayev K andSafarbayeva N2021 E3S Web of Conferences.**264** 02059
12. Abdullayev A,Zhuvanov K and Ruzmetov K 2021 Journal of Physics: Conference Series.**1889(2)** 022121
13. Yuldashev T K, Islomov B I andAbdullaev A A2021 Lobachevskii Journal of Mathematics.**42(3)** pp 663–675
14. Abdullayev A A andErgashev T G2020 Vestnik Tomskogo Gosudarstvennogo Universiteta, Matematika i Mekhanika.**65** pp 5–21
15. VahobovV, Abdullayev A, Kholturayev K, Hidoyatova M andRaxmatullayev A.2020 Journal of Critical Reviews.**7(11)** pp 330–332.
16. Abdullaev A and Hidoyatova M2020 Journal of Critical Reviews**7(11)** pp 337–339.