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, student 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.

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.v.vakhabov2019@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).
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 $y = a + bx$, where:

- $y$—equalized values of the resulting attribute (dependent variable);
- $x$—value of the factor sign (independent variable);
- $a$—the starting point, or the value $y$ at $x = 0$ (it makes no economic sense);
- $b$—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.

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:

$$\sum (y_i - \bar{y}_k)^2 = \min$$

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:

$$\begin{align*}
\sum y &= an + bx, \\
\sum yx &= a \sum x + b \sum x^2
\end{align*}$$
\[ b = \frac{n \sum xy - \sum y \sum x}{n \sum x^2 - \sum x \sum x}; a = \frac{\sum y \sum x^2 - \sum xy \sum x}{n \sum x^2 - \sum x \sum x} \]

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

\[ \hat{y}_x = a + bx + cx^2. \]

\[ \left\{ \begin{array}{l} \sum y = an + b \sum x + c \sum x^2, \\
\sum xy = a \sum x + b \sum x^2 + c \sum x^3, \\
\sum yx^2 = a \sum x^2 + b \sum x^3 + c \sum x^4, \\
\end{array} \right. \]

\[ r = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y}, \]

\[ \bar{xy} = \frac{\sum xy}{n}; \bar{x} = \frac{\sum x}{n}; \bar{y} = \frac{\sum y}{n}; \sigma_x = \sqrt{\frac{\sum x^2}{n} - \bar{x}^2}; \sigma_y = \sqrt{\frac{\sum y^2}{n} - \bar{y}^2}. \]

\[ \sigma^2_{\text{general}} = \sigma^2_{\text{factor}} + \sigma^2_{\text{residual}} \]

\[ r = \frac{\sigma^2_{\text{factor}}}{\sigma^2_{\text{general}}} \]

\[ \sigma^2_{\text{general}} = \sigma^2_{\text{factor}} + \sigma^2_{\text{residual}} \]

\[ r = \sqrt{1 - \frac{\sigma^2_{\text{residual}}}{\sigma^2_{\text{general}}}} \]
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 ± 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 indicators 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>
<thead>
<tr>
<th>$X$ – the yield (centner/ha)</th>
<th>8</th>
<th>11</th>
<th>13</th>
<th>19</th>
<th>21</th>
<th>27</th>
<th>29</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y$ – the prime cost for 1 centner (rouble)</td>
<td>12</td>
<td>8</td>
<td>7</td>
<td>6.3</td>
<td>6</td>
<td>5.8</td>
<td>5.2</td>
</tr>
</tbody>
</table>

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

Fig. 1. The graph shows that, in this case, the connection is close to hyperbole, and the second-order (E3S Web of Conferences 401, 05053 (2023))
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:

\[
\begin{cases}
    a \sum \frac{1}{x^2} + b \sum \frac{1}{x^4} = \sum \frac{y}{x^2} \\
    an + b \sum \frac{1}{x^2} = \sum y
\end{cases}
\]

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

\[
a = \frac{1}{D} \left( \sum y \sum \frac{1}{x^4} - \sum \frac{y}{x^2} \sum \frac{1}{x^2} \right)
\]

\[
b = \frac{1}{D} \left( n \sum \frac{y}{x^2} - \sum y \sum \frac{1}{x^2} \right)
\]

\[
D = n \sum \frac{1}{x^4} - \left( \sum \frac{1}{x^2} \right)^2 \]

Table 2. Calculating a summary table for different \( X \) values

<table>
<thead>
<tr>
<th>( X )</th>
<th>( \frac{1}{x^2} )</th>
<th>( \frac{1}{x^4} )</th>
<th>( y )</th>
<th>( \frac{1}{x^2} )</th>
<th>( \frac{1}{x^4} )</th>
<th>( \bar{y}_x )</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>0.003</td>
<td>1</td>
<td>0.003</td>
<td>0.003</td>
<td>0.12</td>
</tr>
<tr>
<td>9</td>
<td>0.023</td>
<td>0.0002</td>
<td>0.003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.05</td>
</tr>
<tr>
<td>10</td>
<td>0.008</td>
<td>0.00008</td>
<td>0.003</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.03</td>
</tr>
<tr>
<td>11</td>
<td>0.004</td>
<td>0.00004</td>
<td>0.003</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.02</td>
</tr>
<tr>
<td>12</td>
<td>0.002</td>
<td>0.00002</td>
<td>0.003</td>
<td>0.00003</td>
<td>0.00003</td>
<td>0.01</td>
</tr>
</tbody>
</table>

\[
A = \sum y \sum \frac{1}{x^2} - \sum \frac{y}{x^2} \sum \frac{1}{x^2} = 50.6 \cdot 1.4784 - 21.80 \cdot 2.3860 = 22.7922
\]
\[ B = n \sum \frac{y}{x^2} - \sum y \sum \frac{1}{x^2} = 7 \cdot 21.80 - 50.6 \cdot 2.3860 = 318684. \]

\[ a = \frac{A}{D} = \frac{22.77}{4.65} = 4.895 \quad b = \frac{B}{D} = \frac{31.868}{4.656} = 6.8445 \]

\[ \bar{y}_x = 4.9 + \frac{6.8}{x^2} \]

\[ \bar{y}_x = a + \frac{b}{x^2} \]

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

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

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

\[ D_1 = n \sum \frac{1}{x^3} - \left( \sum \frac{1}{x^3} \right)^2 \]
The equations is used:

Sometimes, when studying the relationship between the values of the dependent variable

\[ T \]

The equalizing values of the independent variable X, they will be 

\[ X \]

And so, the empirical regression equation

\[ Y = f + bx + \frac{c}{x} \]

\[ y = f + bx + \frac{c}{x} \]

\[ \sum xy, \sum \frac{1}{x}, \sum \frac{1}{x^2} \]

\[ \sum \frac{1}{x} = \sum y; \]

\[ a \sum x + b \sum x^2 + nc = \sum xy \]

\[ a \sum \frac{1}{x} + bn + c \sum \frac{1}{x^2} = \sum y \]

\[ \sum xy, \sum \frac{1}{x}, \sum \frac{1}{x^2} \]

\[ \sum \frac{1}{x} = \sum y; \]

\[ a \sum x + b \sum x^2 + nc = \sum xy \]

\[ a \sum \frac{1}{x} + bn + c \sum \frac{1}{x^2} = \sum y \]

\[ \sum xy, \sum \frac{1}{x}, \sum \frac{1}{x^2} \]
The connection between the parameters may be described using equation (3). To simplify the computational work, we denote the variables $n, S, \bar{x}, \bar{y}, \bar{X}, \bar{X}^2$ by $a, b, c$ respectively, by $\Sigma$.

Based on the results of this table, we calculate $t$ number of tests ($n$), the error value of the average result ($S$), and the $X, Y$ value. To simplify the expression, $X$ is expressed as a series of natural numbers 1, 2, 3, …

$\sum 45$ by

Table 4.

Table 5.

$$
\begin{align*}
\text{Table 5.} & \\
X & Y & XY & \frac{Y}{X} & X^2 & \frac{1}{X} & \frac{1}{X^2} & \bar{Y}x \\
\hline
\hline
1 & 9 & 1 & 9 & 1 & 9 & 9 & 9 \\
2 & 8 & 1 & 8 & 1 & 8 & 8 & 8 \\
3 & 7 & 1 & 7 & 1 & 7 & 7 & 7 \\
4 & 6 & 1 & 6 & 1 & 6 & 6 & 6 \\
5 & 5 & 1 & 5 & 1 & 5 & 5 & 5 \\
6 & 4 & 1 & 4 & 1 & 4 & 4 & 4 \\
7 & 3 & 1 & 3 & 1 & 3 & 3 & 3 \\
8 & 2 & 1 & 2 & 1 & 2 & 2 & 2 \\
9 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hline
\sum 45 & & & & & & & \\
\end{align*}
$$

Solving this system of normal equations (4):

$$
\begin{align*}
\left\{
9a + 45b + 2.83c &= 17.6 \\
45a + 28b + 9c &= 59.0 \\
2.83a + 9b + 1.54c &= 9.412
\end{align*}
$$

we find:

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

Then the regression equation for the $Y$ by $X$ is expressed as:

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

And the regression equation for the $Y$ by $X$ is:

$$
\bar{Y}_x = a + b x + \frac{c}{x^2}; \quad \bar{Y}_x = a + b x + \frac{c}{x^3}
$$
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 $Y$ decreases rapidly, then it is convenient to use the third-order hyperbola equation to flatten the empirical series.

References


2. Gmurman V Ye 1977 Probability theory and mathematical statistics (Moscow: High school) p 480


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 L 1957 Mathematics for economists (Moscow) p 267


9. Fedoseev V V 2000 Economics-mathematical methods and applied models (Moscow) p 345


11. Abdullayev A, Kholturyev K and Safarbayeva N 2021 E3S Web of Conferences. 022121


