General Concept About Correlation Relationships.

Istamjon Karshiev
Tashkent University of Applied Sciences, Gavhar Str. 1, Tashkent 100149, Uzbekistan
istamkarshiev114@gmail.com

https://doi.org/10.5281/zenodo.10471195

Keywords: correlation, determination coefficient, regression equation, Fisher's criterion, determination coefficient, multiple correlation, Kramer, constant coefficients, Student criterion.

Abstract: This annotation provides a comprehensive insight into the general concept of correlation relationships. It explores the fundamental principles and applications of correlations within various fields, emphasizing their significance in statistical analysis, predictive modeling, and decision-making processes. The annotation delves into the types of correlations, their mathematical representation, and the interpretation of correlation coefficients. Additionally, it discusses the strengths and limitations of correlational studies, shedding light on their role in establishing connections between variables and elucidating associations within datasets. The annotation aims to offer a foundational understanding of correlation relationships, serving as a valuable resource for researchers, analysts, and students exploring this critical statistical concept.

1 INTRODUCTION

Topic objective. To illustrate the relationship (dependence) between economic indicators.

In economics, technology, and other fields, it is appropriate to study the relationships between various numerical and qualitative indicators. Among the indicators, there are two types of relationships - functional and correlation (or statistical) relationships.

In functional relationships, for any arbitrary value of one variable, there is a specific value of another variable that corresponds to it. Such relationships are specifically studied in subjects such as mathematics, physics, and chemistry.

Topic objective. To illustrate the relationship (dependence) between economic indicators.

In economics, technology, and other fields, it is appropriate to study the relationships between various numerical and qualitative indicators. Among the indicators, there are two types of relationships - functional and correlation (or statistical) relationships.

In functional relationships, for any arbitrary value of one variable, there is a specific value of another variable that corresponds to it. Such relationships are specifically studied in subjects such as mathematics, physics, and chemistry.

If there is such a relationship (correlation) between two random variables, and the average value of one variable changes in a regular manner with the change in the average value of the other variable, then such a relationship is called a statistical or correlation relationship. Correlation models are very important models in the group of economic-statistical models.

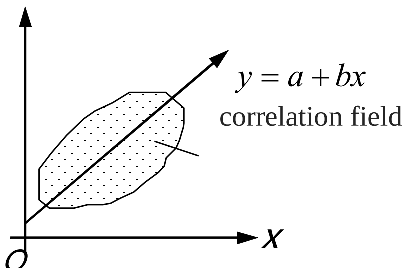
In the simplest case, the relationship between two indicators or variables is tested. In this case, it is called a simple correlation. If there is a relationship between two or more indicators, then it is called a multiple correlation.

If an indicator is related to other indicators, it is called a dependent indicator and is identified with it, as well as with the variables that are dependent on it, such as argument or factor variables or simply factors.

In this case, the correlation model is written as follows:

$$\widetilde{y}_i = a + bx_i$$

The regression equation is a formula that expresses the relationship between the dependent variable, which is found through testing, the independent variables obtained from experiments, and the constant coefficients. They have the following economic meanings: a represents the significance of the unaccounted factors, b indicates how much y will increase when x increases by one unit. The correlation is represented graphically as follows. [1]



(1)

Creating correlation models consists of several stages:

1) Formulating the problem;

2) Collecting statistical data;

3) Determining the form of the regression equation;

4) Finding the strength of the relationship;

5) Finding finite values for the parameters in the regression equation;

6) Applying the obtained result to economics.

Let's say we have a regression equation with an unknown variable $\tilde{y}_i = a + bx_i$ where the size of the sample is N. To find the equation a and b its coefficients, we will use the method of least squares:

$$S = \sum_{i=1}^{N}\left(y_i - \tilde{y}_i\right)^2 = \min \qquad (2)$$

Here we y_i - present the positive indicators that can be obtained from experience: (1) and (2).

$$S = \sum_{i=1}^{N}\left(y_i - a - bx_i\right)^2 = \min$$

We obtain special properties with respect to the normal distribution, i.e.

$$\left.\begin{aligned} \frac{\partial S}{\partial a} &= -2\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0 \\ \frac{\partial S}{\partial b} &= -2x_i\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0 \end{aligned}\right\} \qquad (3)$$

(3) Let's multiply the base of the triangle by (-2) and divide the system into both sides.

$$\left.\begin{aligned} &\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0 \\ &x_i\sum_{i=1}^{N}\left(y_i - a - bx_i\right) = 0 \end{aligned}\right\} \quad (4)$$

(4) We open the locks and simplify.

$$\left.\begin{aligned} Na + b\sum_{i=1}^{N}x_i &= \sum_{i=1}^{N}y_i \\ a\sum_{i=1}^{N}x_i + b\sum_{i=1}^{N}x_i^2 &= \sum_{i=1}^{N}x_i y_i \end{aligned}\right\} \quad (5)$$

(5) Kramer left with a sad face.

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

$$b = \frac{\begin{vmatrix} N & \sum_{i=1}^{N}y_i \\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i y_i \end{vmatrix}}{\begin{vmatrix} N & \sum_{i=1}^{N}x_i \\ \sum_{i=1}^{N}x_i & \sum_{i=1}^{N}x_i^2 \end{vmatrix}}$$ [2]

The correlation coefficient is used to determine the strength of the relationship.

$$r_{yx} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{\left(N-1\right)\ \delta_x\ \delta_y}$$

$$\delta_x = \sqrt{\frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{\left(N-1\right)}}$$

$$\delta_y = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{\left(N-1\right)}}$$

If r_{yx} is close to 1, it is called a strong connection, otherwise it is called a weak connection, and it varies in the range of $0 < r_{yx} \le 1$. [3]

Analyse the correlation relationship between the cost incurred by the British government for road construction from 2013 to 2023, denoted as Y (in million pounds sterling), and the time. The results required to find the coefficients of the regression equation based on the economic information provided in the table are presented.

Year	t	t^2	y_i	y_i^2	$t\ y_i$
2013	1	1	560	313600	560
2014	2	4	608	369664	1216
2015	3	9	685	469225	2055
2016	4	16	807	651249	3228
2017	5	25	839	703921	4195
2018	6	36	914	935396	5484
2019	7	49	1100	1210000	7700
2020	8	64	1196	1430416	9568
2021	9	81	1490	2247001	13491
2022	10	100	1544	2477476	15740
2023	11	121	1513	2289169	16643
Σ	66	506	11295	12997117	79880

$$b=\frac{11\cdot 79880}{11\cdot 506-66^2}=\frac{878680-745470}{5566-4356}$$
$$=\frac{133210}{1210}=110{,}09$$

$$a=\frac{11295}{11}-\frac{110{,}09\cdot 66}{11}=1026{,}81-660{,}54=366{,}27$$

The equation of regression in the result.

$$\tilde{y}_i=366{,}27+110{,}09\cdot t$$

It will be seen in the future. This discovered regression equation provides the opportunity to predict for the next year, i.e., when t=12.

$$\tilde{y}_{12}=366{,}27+110{,}09\cdot 12=1687{,}35$$

The correlation coefficient of $r=0{,}95$ indicates that there is a strong positive correlation between the variables. [4]

The Student criterion is used to determine the consistency of coefficients.

$$t_j=\frac{|b_j|}{S_{b_j}}$$

In this equation, b_j is the coefficient of the b_j -regression; S_{b_j} is the average quadratic deviation of the coefficient j .

$$S_{b_j}=\sqrt{\sum_{i=1}^{N}\left(\frac{\partial b_j}{\partial y_j}\right)^2\cdot S_i^2}$$

$$S_i^2=\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}_i\right)^2}{N-1}$$

We use the Fisher criterion to determine the stability of the regression equation.

$$F=\frac{S_y^2}{\bar{S}^2}$$

$$S_y^2=\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2}{N-1} \qquad \bar{S}_{\tilde{y}}^2=\frac{\sum_{i=1}^{N}(\tilde{y}_i-\bar{\tilde{y}})^2}{N-1}$$

When these criteria are taken into account, the larger the result is compared to the result in the table, the more stable the regression equation becomes. [5]

2. Multiple factor correlation relationship.

When testing practical problems, the correlation relationship is usually dependent on factors that have a significant impact, and in this case, the regression equation takes the following form: In this case, the regression equation is: $\tilde{y}_i=b_0+b_1x_{1i}+b_2x_{2i}+\ldots+b_kx_k$

Let's write the given statistical data in the form of a table: (1)

№	x_{1i}	x_{2i}	x_{3i}	...	x_{ki}	y_i
1	x_{11}	x_{21}	x_{31}	...	x_{k1}	y
2	x_{12}	x_{22}	x_{32}	...	x_{k2}	y_2
3	x_{13}	x_{23}	x_{33}	...	x_{k3}	y_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	x_{1N}	x_{2N}	x_{3N}	...	x_{kN}	y_N

Let's convert the measurement from the old unit to the new unit using the following formulas:

$$y_i^0 = \frac{y_i - \bar{y}}{Sy} \qquad x_{ji}^0 = \frac{x_{ji} - \bar{x}_j}{Sx_j} \qquad \begin{matrix} i = 1,2,\ldots,N \\ j = 1,2,\ldots,k \end{matrix}$$

Here, y_i^0 and x_{ji}^0 are the normalized values of the factors; $\bar{y}$ and $\bar{x}_j$ are the average values of the factors; S_y and $S_{x_j}$are the average quadratic deviations of the factors.

$$S_y = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y}_i)^2}{N-1}} \qquad S_{x_j} = \sqrt{\frac{\sum_{i=1}^N \left(x_{ji} - \bar{x}_j\right)^2}{N-1}} \quad [6]$$

We will present statistical data in the 2nd table with new measurements.

№	x_1^0	x_1^0	x_1^0	...	y_i
1	x_{11}^0	x_{21}^0	x_{31}^0	...	y_i^0
2	x_{12}^0	x_{22}^0	x_{32}^0	...	y_2^0
3	x_{13}^0	x_{23}^0	x_{33}^0	...	y_3^0
⋮	⋮	⋮	⋮	⋮	⋮
N	x_{1N}^0	x_{2N}^0	x_{3N}^0	...	y_N^0

On a new scale

$$\bar{x}_j^0 = 0, \quad \bar{y}_0 = 0$$

$$S_{x_j^0} = 1, \quad S_{y^0} = 1 \quad (3)$$

The coefficient of determination is given by the following formula:

$$\left.\begin{aligned} r_{y^0 x_j^0}^* &= \frac{1}{N-1} \sum_{i=1}^N y_i^0 x_{ji}^0 \\ r_{x_i^0 x_m^0}^* &= \frac{1}{N-1} \sum_{i=1}^N x_{ji}^0 x_{mi}^0 \end{aligned}\right\}$$

$$l > m \quad l, m = 1,2,\ldots,k \quad (4)$$

In general, the error term in the regression equation is not independent and follows the following pattern:

$$\tilde{y}^0 = a_1 x_1^0 + a_2 x_2^0 + \ldots + a_k x_k^0 \quad (5)$$

(5) The coefficients are found from this condition:

$$S = \sum_{i=1}^N (y_i^0 - \tilde{y}_i^0)^2 = min$$

The minimum condition of the function is found in only one variable.

$$\frac{\partial S}{\partial a_1} = 0 \qquad \frac{\partial S}{\partial a_2} = 0 \ldots \frac{\partial S}{\partial a_k} = 0 \qquad (6)$$

$$a_1 \sum_{i=1}^N (x_{1i}^0)^2 + a_2 \sum_{i=1}^N x_{1i}^0 x_{2i}^0 + \ldots + a_k \sum_{i=1}^N x_{1i}^0 x_{ki}^0 = \sum_{i=1}^N x_{1i}^0 y_i^0$$
$$a_1 \sum_{i=1}^N x_{2i}^0 x_{1i}^0 + a_2 \sum_{i=1}^N (x_{2i}^0) 0{,}2 + \ldots + a_k \sum_{i=1}^N x_{2i}^0 x_{ki}^0 = \sum_{i=1}^N x_{2i}^0 y_i^0$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$a_1 \sum_{i=1}^N x_{ki}^0 x_{1i}^0 + a_2 \sum_{i=1}^N x_{ki}^0 x_{2i}^0 + \ldots + a_k \sum_{i=1}^N (x_{ki}^0)^2 = \sum_{i=1}^N x_{ki}^0 y_i^0$$

(7)

We increase the left and right sides of the system by $\frac{1}{N-1}$. As a result, the determination coefficient r^* is mainly produced in front of the a_j coefficients, which is (7).

$$\frac{1}{N-1} \sum_{i=1}^N \left(x_{ji}^0\right)^2 = S_{x_j^0}^2 = 1 \quad [7]$$

Let's create a normal equations system:

$$\left.\begin{aligned} &a_1 + a_2 r_{x_1 x_2}^* + a_3 r_{x_1 x_3}^* + \ldots + a_k r_{x_1 x_k}^* = r_{y x_1}^* \\ &a_1 r_{x_2 x_1}^* + a_2 + a_3 r_{x_2 x_3}^* + \ldots + a_k r_{x_2 x_k}^* = r_{y x_2}^* \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &a_1 r_{x_k x_1}^* + a_2 r_{x_k x_2}^* + a_3 r_{x_k x_3}^* \ldots + a_k = r_{y x_k}^* \end{aligned}\right\}$$

If we calculate the correlation coefficients in $r_{x_l x_m}^* = r_{x_m x_i}^*$, they are found to be zero.

(8) It is possible to calculate the correlation coefficient R for a large number of factors.

$$R = \sqrt{a_1 r_{y x_1}^* + a_2 r_{y x_2}^* + \ldots + a_k r_{y x_k}^*}$$

The correlation coefficient R varies in the following range:

$$0 \leq R \leq 1$$

If the value of R is close to 1, the correlation relationship is considered strong or stable. To use (5) in practice, we need to convert it to the natural scale using these formulas:

$$b_j = a_j \frac{S_y}{S_{x_y}} \qquad \begin{matrix} j = 1,2,\ldots,k \\ j \neq 0 \end{matrix}$$

$$b_0 = \bar{y} - \sum_{j=1}^R b_j \bar{x}_j$$

An algorithm has been written in "Pascal" language for the mentioned algorithm, making it convenient to solve practical problems. [8]

Years	Indicators			
	Labor productivity	Armed forces fund x_1	Specialization level x_2	Workload x_3

1	100	100	100	100
2	109	108	140	109
3	119	130	104	123
4	132	135	99	139
5	139	212	97	153
6	148	152	102	157
7	156	158	120	164
8	165	265	147	171
9	173	288	151	179
10	188	305	155	195
11	192	325	171	204
12	197	350	199	215
13	202	360	226	228
14	206	400	226	236
15	211	392	238	256

CONCLUSIONS

The regression equation for this issue

$$\tilde{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{3i} \quad (1)$$

Need to find Z from (1) $j = 0,1,2,3$

We use the formula of correlation to find unknown b_j, with $j = \overline{0,4}$ coefficients, according to the instructions of the program: refer to the results obtained from the computer. $b_0 = 43{,}9$

$b_1 = 0{,}157 \quad b_2 = -0{,}026 \quad b_3 = 0{,}467$.

If we substitute the obtained results into (1)

$$\tilde{y}_i = 43{,}9 + 0{,}157 x_{1i} - 0{,}026 x_{2i} + 0{,}467 x_{3i}$$

The stability of the relationship found in (2) is indicated by the correlation coefficient.

$R = 0{,}985$

Fisher's criterion $F = 182{,}3$ $(F_{jadv} = 6{,}54)$ When F=1, it indicates the correctness of equation (2).

ACKNOWLEDGMENTS

The author thanks Kayumov E. and Ataniyazova M., professors and teachers of Tashkent University of Applied Sciences, for their scientific and practical help in writing this article.

REFERENCES

1. Т.Шодиев. Эконометрика. Тошкент 1999.
2. Т.Шодиев, А.Кучкоров, У.Мирзапов. Ишлаб чиšаришни режалаштиришда математик усуллар. Тошкент. Ўзбекистон. 1995.
3. К.Д.Льюнс. Методы прогнозирования экономических показателей. Москва. 1986.
4. Л.Плариронов, Т.И.Юрченко, Л.Л.Новосёлов. Экономико-математические методы в планировании.
5. С.С.Резниченко, М.П.Подольский, А.А.Ашихмин. Экономико-математические методы и моделирование в планировании и управлении горным производством. Москва. Недра. 1991.
6. Л.Э.Хазанова. Математическое моделирование в экономике. М: БЕК, 1998.
7. В.А.Колемаев. Математическая экономика. М.: ЮНИТИ.1998.
8. А.А.Горчаков, И.В.Орлова. Компьютерные экономико-математические модели. М.:ЮНИТИ. 1995.