Correlation-regression analysis in Excel when solving problems

Smirnova Anna Sergeevna

Candidate of Pedagogic Sciences, Associate Professor

Sholem Aleichem Priamursky State University

Abstract. The article discusses the application of the methods of correlation and regression analysis to the description of the relationship between the area of a store and the volume of its annual sales; a linear model has been built that makes it possible to predict store income, determined by its area.

Keywords: statistics, correlation-regression analysis, Pearson's test, paired linear regression, coefficient of determination.

Statistical methods of information processing are used in physics, chemistry, biology, economics, psychology and other sciences. Analysis and forecast of socio-economic phenomena, planning of production of goods and services is based on statistical methods [2]. To organize and present statistical data, statistical series and graphs are used.

For the convenience of calculations, spreadsheets are used. The main advantage and difference of spreadsheets is the ease of use of data processing tools. Data processing tools in their capabilities can be compared with databases; working with them does not require special training in programming from the researcher. You can enter any information into tables: text, numbers, dates and times, formulas, pictures, diagrams, graphs. All entered information can be processed using special functions [3]. MS Excel for Windows has a powerful mathematical statistics tool that allows you to do statistical modeling.

The *r*-Pearson correlation coefficient is used to study the relationship of two metric variables measured on the same sample. The coefficient characterizes the presence of only a linear relationship between the features, usually denoted by the symbols *x* and *y*.

Problem. To assess the relationship between store size (in square feet) and annual sales, consider a sample of 14 stores. The question is - is there a relationship between the area of the store and the volume of its annual sales [1]?

Solution. x - store area (thousand square feet), y - annual sales (million dollars).

Hypotheses:

H_0: The correlation between variables x and y does not differ from zero.

H_1: The correlation between variables x and y is significantly different from zero.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14

x	1.7	1.6	2.8	5.6	1.3	2.2	1.3	1.1	3.2	1.5	5.2	4.6	5.8	3
y	3.7	3.9	6.7	9.5	3.4	5.6	3.7	2.7	5.5	2.9	10.7	7.6	11.8	4.1

To advance a hypothesis, Insert / Diagrams are used: Point. Let's build a scatter diagram (fig. 1). It can be assumed that there is a linear positive correlation, therefore, the Pearson correlation coefficient can be applied.

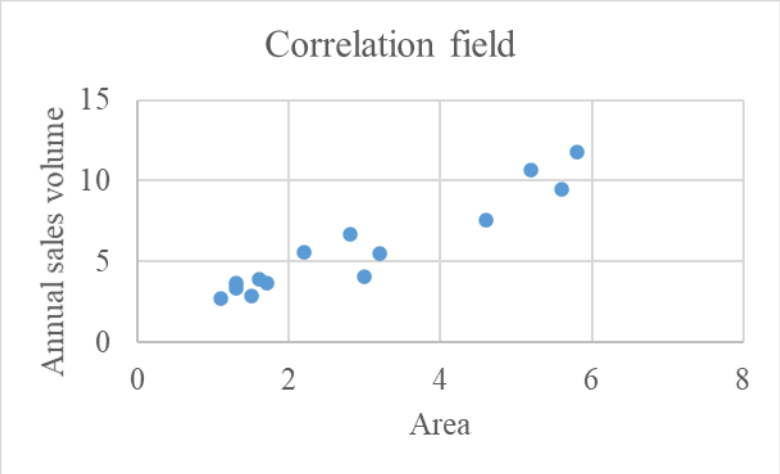

Figure 1 – Scatter diagram

Calculation of the Pearson correlation coefficient using the CORREL statistical function (fig. 2)

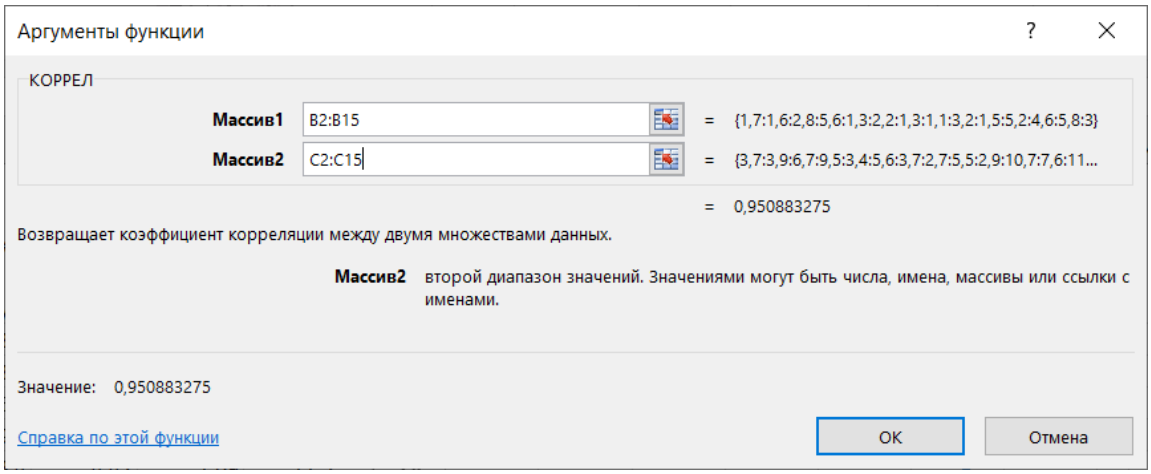

Figure 2 – Calculating the Pearson correlation coefficient

We determine the critical values and the construction of significance from the table (fig. 3).

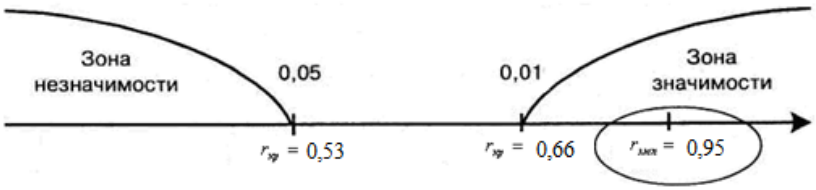

Figure 3 – "Significance axis" built for Pearson's criterion

Value $r_{эмп}$=0.95 falls into the zone of significance, an alternative hypothesis is accepted at a significance level of 1%, i.e. There is a strong positive correlation between the area of a store and its annual sales. The resulting directly proportional relationship suggests that the larger the store area, the greater the volume of its annual sales, and vice versa.

The study of the problem can be continued using regression analysis. The main task of which is to find the coefficients of a_0, a_1 and the regression equations of $y = a_0 + a_1 \cdot x$. The regression coefficient a_1 shows how much, on average, the value of one variable changes when the measure of another variable changes by one unit.

Continuation of the task. Determine if a 1 square foot increase in store floor space will increase the store's annual sales.

Decision. We use paired linear regression analysis to answer the question. Into the constructed correlation floor, we introduce the trend line and the coefficient of determination (fig. 4 and 5).

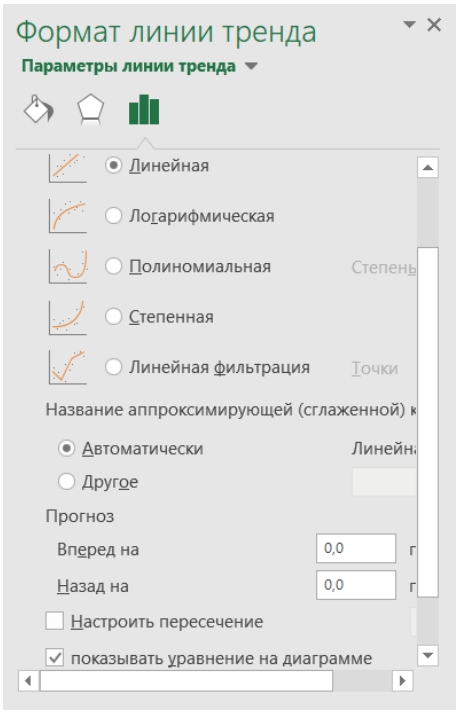

Figure 4 – Trendline format

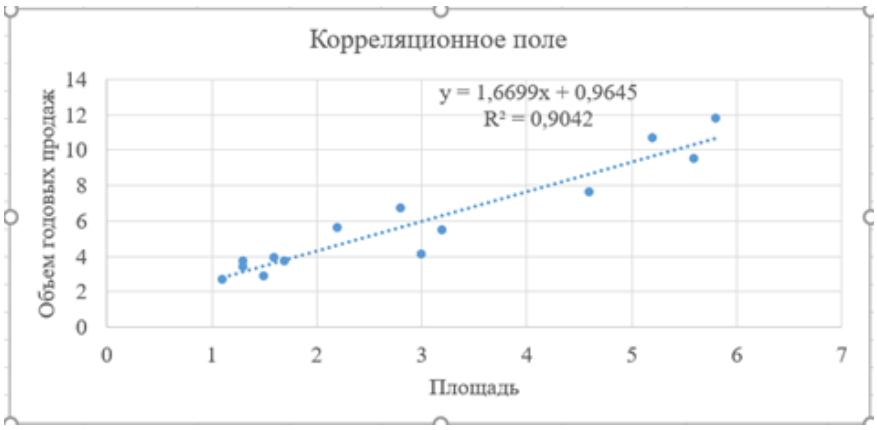

Figure 5 – Regression equation and coefficient of determination

Regression equation $y = 0{,}96 + 1{,}67 \cdot x$

The calculated slope a_1= +1.67 means that if the store is increased by one square foot, then the annual sales will increase by 1.67 thousand dollars. Therefore, the slope is the proportion of the annual sales as a function of store size. The shift a_0= +0.9645 million dollars) determines the average value of the variable y at x = 0. But the store area cannot be equal to zero, which means that the shift can be considered a share of the annual income, depending on other factors [1].

The coefficient of determination R2 shows what proportion (in%) of the change in the effective attribute is caused by the change in the factor attribute x. In the problem, R^2=0.9 means that 90% of the differences in annual sales of stores are explained by the difference in their total area, and 10% - by other unaccounted factors.

Models with a coefficient of determination above 80% can be considered quite good. Therefore, you can use a linear regression model to predict annual store sales based on store size. Let the store area be 4,000 square feet. Let's predict the average annual sales by substituting the value x = 4 (thousand square feet) in the linear regression equation of $y = 0{,}96 + 1{,}67 \cdot x = 0{,}96 + 1{,}67 \cdot 4 = 7{,}64$ million dollars. So, the projected average annual sales in a store with an area of 4000 sq. feet is 7.640 million dollars [1].

Thus, the methods of correlation and regression analysis can be used to identify links between several factors of financial and economic activity and to assess the closeness of interdependence of the factors selected for analysis.

References

1. Simple linear regression – URL: https://baguzin.ru/wp/prostaya-linejnaya-regressiya/

2. Statistics as an independent social science – URL: http://studies.in.ua/pravovaya-statistika-seminar/2035-statistika-kak-samostoyatelnaya-obschestvennaya-nauka.html

3. MICROSOFT EXCEL – URL: https://portal.tpu.ru/SHARED/m/MARTYNOVYAA/study_work/ktit/labs/%D0%A2%D0%B5%D0%BE%D1%80%D0%B8%D1%8F%20Excel_0.pdf