www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC Gestão, Inovação e Tecnologias

Correlation and Regression Analysis of Economic Problems

Lyudmila Valentinovna Bolshakova[1]; Alexander Nikolaevich Litvinenko[2];
Inna Kazimirovna Sidenko[3]; Anatoly Nikolaevich Ivanov[4]; Grigory Leonidovich Shidlovsky[5];
Boris Semyonovich Limonov[6]; Farid Abdulalievich Dali[7]
[1,2]Saint Petersburg University of the Ministry of Internal Affairs of Russia, Russia.
[3]Russian State Hydrometeorological University, Russia.
[4,5,6,7]Saint Petersburg University of the State Fire Service of the EMERCOM of Russia, Russia.

Abstract
The article considers the problem related to the correlation and regression analysis. The use of correlation and regression analysis methods is shown in the specific example illustrating the study of a particular economic situation. Special attention is paid to the use of the Excel Analysis Package, which significantly simplifies the solution of the problem from a computational standpoint.

Key-words: Correlation and Regression Analysis, Multiple Linear Regression Equation, Correlation and Determination Coefficients, Adequacy.

1. Introduction

When studying various processes and phenomena of scientific interest, quite often there is a need to clarify the features and properties of various factors that influence these processes and phenomena. In the presence of a large statistical material, multidimensional mathematical and statistical methods can be successfully used to study these factors [3], which also include correlation and regression analysis methods. Using these methods allows obtaining conclusions and forecasts that can confirm or refute particular results of scientific research.

The scope of application of correlation-regression analysis is quite wide and diverse, due to the universality and variety of available methods. However, from a scientific standpoint, the most interesting results of employing correlation and regression analysis can be obtained in the economic realm. For example, the study of the impact of price, supply, advertising costs, etc. on the demand for

a certain product; the analysis of the dependence of GDP growth on the factors of the innovative economy [10] or changes in the structure of employment in the economy [11]; the study of economic growth in the region [4]; or the study of inflation using econometric analysis [5]. Quite interesting correlation and regression studies can be carried out in the banking sector [2, 9].

The methods of correlation and regression analysis allow determining the existence and strength of the relationship of factors based on sample statistical data, as well as obtaining the type of dependence of one factor on another or several factors.

Correlation and regression analysis has the following advantages:

- the possibility of a comprehensive study of various relationships between factors;
- obtaining an assessment of the resulting factor behavior, including its possible predicted values that are adequate to reality.

The disadvantages of correlation and regression analysis methods include:

- cumbersome calculations;
- a significant impact of the sample size and composition on the results obtained;
- the need to fulfill the prerequisites of the least-squares method.

The first drawback can be almost eliminated by employing the Excel Analysis Package for conducting correlation and regression analysis. The elimination of the second disadvantage is directly related to the capabilities of the researcher. And, finally, for correctly obtained samples of a large volume, most often, the prerequisites of the least-squares method are fulfilled [7].

The purpose of the present article is to analyze the possibilities of using correlation and regression analysis, to detail the methodology of its use, and the correctness of the interpretation of the results when conducting scientific research.

2. Methods

Research design

To achieve this goal, a research strategy was chosen based on a mixed approach of qualitative and quantitative methods of data collection and analysis.

Due to the limited resources at our disposal, but observing the most important principles of scientific research (obtaining more reasonable and reliable results), a CASE (computer-aided software engineering) technology for analyzing the management system using correlation and regression

analysis was chosen. It is illustrated in the example of analyzing the relationship between the volumes of the company's products sold with the presence of counterfeit analogs in a competitive market.

In the course of conducting a general correlation and regression study, two problems were solved, namely, a correlation and regression analysis of the sample population was carried out, then the results of the sample analysis were distributed to the entire general population (to confirm the adequacy of the results of the sample study).

The sequence and main tasks of the correlation and regression study for the linear case were as follows:

1. Problem statement.
2. Constructing and studying a simple linear regression equation.
3. Checking the regression model for adequacy.
4. Determining the strength of the dependence between the signs.
5. Getting conclusions and forecasts.

Case study

An enterprise that produces a certain type of product was selected for the study. The company's executives drew attention to the fact that in recent months, the volume of products sold has begun to fall sharply. It was found out that the decline in sales volume was due to the appearance of fake, cheaper products of the same type on the market. The management and marketing department of the enterprise collected data on the volumes of counterfeit products and the volumes of products sold for the past months of the current year presented in Table 1.

Table 1:Different production volumes by month

Volume, thousand pcs.	Jan	Feb	March	Apr	May	June	July	Aug	Sept	Oct
Released products	32	28	30	27	29	26	27	32	30	28
Products sold	27	24	25	20.9	21.5	19.8	19	16.9	17.8	16.4
Counterfeit products	3.67	4.36	5	5.56	6	6.7	7.77	8.78	9.36	10.53

The task was set to predict the volume of products that can be sold in December under the same conditions if the company became aware of the estimated volume of counterfeit products

expected in December, equal to 10 thousand pieces. In December, the company expects to release 30 thousand pieces of products.

To solve this problem, a correlation and regression analysis of the presented data was conducted using an Analysis Package (Solution Analysis). To do this, data was entered in Excel, choosing Y – the volume of products sold as the resulting attribute. Factor signs were the volume of counterfeit products X1 and the volume of products released by the enterprise X2.

Description of the calculation

Data were entered by columns and presented in the form of Table 2.

Table 2: Task data

Month	The volume of counterfeit products X_1	The volume of released products X_2	The volume of products sold Y
January	3.67	32	27
February	4.36	28	24
March	5	30	25
April	5.5	27	20.9
May	6	29	21,5
June	6.7	26	19.8
July	7.77	27	19
August	8.78	32	16.9
September	9.36	30	17.8
October	10.53	28	16.4

Next, the DATA button was selected in the mainline, and then the Analysis Package (Solution Analysis). In the Analysis Tools window, the Regression line was selected, and the appeared Table was filled in the following way.

Input interval Y: the cells of the last column were assigned the name of the column: Y – the volume of products sold.

Input interval X: the cells of the second and third columns were respectively assigned the names: X1 – the volume of counterfeit products, and X2 – the volume of manufactured products.

Tags: checking the box, since the data were entered with the column names, otherwise the box remains empty.

Constant is zero: the cell remains empty since it is not required that the regression line passes through the origin.

Reliability level: by default, it is assumed that the reliability of the results obtained is 95%.

In the Output interval, the cell from which the presentation of the results begins was registered.

Let's register the remaining cells (without graphs).

After clicking the OK button, the RESULTS are displayed in the form of several Tables.

Below, the semantic content of the obtained results in considered more in detail with the above-specified data. The sequence of Tables will be selected according to the above-listed tasks.

3. Results and Discussion

Construction and investigation of a sample linear regression equation

The sampling linear regression equation for the problem under consideration has the form:

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2,$$

where the sampling coefficients b0, b1, b2 are estimates (approximate values) of the corresponding coefficients of the regression equation characterizing the entire general population.

To determine the coefficients of the multiple linear regression equation, five columns of Table 3 were selected.

Table 3: Coefficients of the regression equation

	Coefficients	*Standard error*	*t-statistics*	*P-value*
Y-crossing	22.32028219	5.644238006	3.954525334	0.0055
x1	-1.458365988	0.171535778	-8.501818114	6,17E-05
x2	0.289912818	0.187842424	1.543383075	0.166647

The second column of Table 3 shows the coefficients of the regression equation. After rounding, the following equation was obtained:

$$\hat{y} = 22.3202 - 1.4584 \cdot x_1 + 0.2899 \cdot x_2.$$

By the values of the regression coefficients, it is possible to determine the contribution of each factor to the change in the values of the resulting feature Y. The coefficient for variable xi shows how much the value of feature Y will change on average if the value of feature xi is increased by one unit.

Therefore, two statements are valid for the concerned problem:

If the volume of counterfeit products increases by one thousand, the volume of products sold will decrease by 1.4584 thousand provided an unchanged volume of the products produced.

If the volume of manufactured products increases by one thousand, the volume of sold products will increase by 0.2899 thousand provided an unchanged volume of the counterfeit products.

To confirm the results obtained, the model was checked for adequacy.

Checking the model for the adequacy

The regression equation was built based on sample data, i.e. using only a part of the general population. Naturally, the question arises as to whether this equation, as well as the model based on this equation, is adequate to reality. The adequacy of the model was checked using the average approximation error.

The average approximation error characterizes the degree of proximity of the sample values of the attribute Y and the corresponding values obtained by the regression equation. This error is determined by the formula:

$$A = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \cdot 100\%.$$

If this error is within the range of 4 to 12%, then the regression model can be considered adequate to reality.

The analysis package does not give the value of this error, but it gives Table 4, whose elements significantly simplify the task of finding the sought error.

Table 4: Determination of residuals

Observation	*The predicted value of Y*	*Residuals*	*Standardized residuals*
1	26.2452892	0.754710798	0.732322855
2	24.0793654	-0.079365396	-0.07701108
3	23.7258368	1.274163199	1.236366081
4	22.12691535	-1.226915351	-1.190519806
5	21.97755799	-0.477557995	-0.463391586
6	20.08696335	-0.286963348	-0.27845079
7	18.81642456	0.183575441	0.17812981
8	18.793039	-1.893039004	-1.836883387
9	17.36736109	0.432638906	0.41980499
10	15.08124725	1.318752749	1.279632914

The second column of Table 4 shows the values of the regression function for the given values of factor features. The third column contains found residuals. It's not hard to notice that these values correspond to those of the numerator of the terms in the formula for the average approximation error. As a result, one gets A = 0.040402 or A = 4.04%, which indicates that the model is adequate.

The last column of Table 4 shows the standardized residuals.

It should be noted that in some cases, the average approximation error gives incorrect information about the adequacy of the model [1]. Therefore, it is recommended to use other methods of checking the adequacy of the model, for example, checking the corresponding statistical hypotheses about significance. When checking the statistical hypothesis about the significance of the equation in general and statistical hypotheses about the significance of the coefficients of the regression equation, the observed values of the criterion found by specific formulas using sampling data are compared with the critical values of the same criterion found by distribution tables. If the observed value (or the modulus of the observed value) is greater than the critical one, then it is assumed that the equation is significant (or the coefficient is significant), i.e. the equation is adequate to reality and can be used for further study, including forecasting. Otherwise, the adequacy of the model remains in question. In this case, it is recommended to re-sample, increase the sample scope, etc., and perhaps even revise the factor features.

The significance of the regression equation, in general, can be checked using the Fisher criterion. To do this, it is necessary to compare two values Fn and Fcr. These values can be determined using Table 5.

Table 5: Variance analysis

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	105.0623082	52.53115408	38.4695	0.000167483
Residual	7	9.558691842	1.365527406		
Total	9	114.621			

The second column of Table 5 defines the number of degrees of freedom df for the regression and the residual.

For regression, df is equal to the number of features m, equal in our case to 2. For residuals, df = $n - (m+1) = 10 - 3 = 7$. These numbers are used to find Fcr. The value of Fn is presented in the fifth column of the Table, i.e. Fn = 38.4695. The value of Fcr can be found using the Fischer-Snedekor

Distribution Table or using the function F.OBR.PH: Fcr = 4.7374. Since Fn > Fcr, the regression equation, in general, can be considered significant and used for forecasting.

Similarly, one can check the significance of the regression coefficients according to the Student's criterion, using the elements of column 4 of Table 3 as a sample value. The third column of Table 3 shows the values of standard errors for each regression coefficient. The absolute values of errors of the coefficients b0 and b1 are significantly less than their actual values, so there is reason to believe that these coefficients differ significantly from zero.

After constructing and studying the sample regression equation, the problem arises of determining the strength of the relationship between the features.

Determining the strength of the relationship between the features

Three types of correlation coefficients can be used to identify the strength of the relationship in the case of multiple regression. The multiple correlation coefficient shows the strength of the impact of all factor features, included in the model, on the resulting feature. The matching correlation coefficient shows the strength of the dependence between two features against the background of the impact of all the others. A partial coefficient of correlation shows the strength of the dependence between two features, excluding the influence of all other features included in the model.

Consider Table 6.

Table 6: Regression statistics

Regression statistics	
Multiple correlation coefficient R	0.957395479
R-square	0.916606103
Normalized R-square	0.892779276
Standard error	1.168557832
Observation	10

The second line of Table 6 shows the multiple correlation coefficient Ry ≈ 0.957395. The coefficient value is quite close to unity, therefore, there is a very strong linear relationship between the resulting feature Y and the two factorials X1 and X2. In other words, the strong influence of the volumes of both counterfeit products and manufactured products on the change in the volume of products sold has been proved.

The determination coefficient R_Y^2, presented in the third line, shows that almost 92% of the change in the volume of products sold is due to the joint influence of the volumes of counterfeit products and manufactured products.

The fourth line indicates the so-called normalized or corrected determination coefficient R2corr. Its introduction is related to the following. The determination coefficient R_Y^2 has one significant drawback. As a rule, its value increases when other factor features are added, even if they do not have a strong impact on the resulting feature. Therefore, this coefficient is slightly corrected. For the problem under consideration, the usual coefficient of determination remains, since the difference between it and the corrected coefficient is quite small (slightly more than 0.02).

To determine the matching coefficients, we go back to the Analysis Tools window and select the Correlation line. After entering the data, we get Table 7, which presents the matching correlation coefficients, i.e. the correlation coefficients between two different features, provided that the third feature also influences.

Table 7: Correlation coefficients

	x1	*x2*	*y*
x1	1		
x2	-0.07167	1	
y	-0.94246	0.235571	1

The following conclusions can be drawn based on Table 7.

The matching correlation coefficient is rYX1 = –0.94246, therefore, the linear relationship between the volumes of counterfeit and sold products is very strong and inverse, i.e. with an increase in the volume of counterfeit products, the volume of sold products decreases.

The matching correlation coefficient rYX2 = 0.235571 shows that the relationship between the volumes of manufactured and sold products, although direct, is weak. This may be influenced by the volume of counterfeit products. Whether this is correct or not will be shown by the partial correlation coefficients. Finally, the matching correlation coefficient rX1X2 = –0.07167 shows the absence of a linear relationship between the factor features, which allows concluding that there is no multicollinearity (the coefficient is less than 0.5). Thus, the results indicating the good quality of the constructed model are confirmed.

The partial correlation coefficients are not calculated in the Analysis Package, and therefore they need to be found using the appropriate formulas. If the regression model contains only two factor features X1 and X2, then the partial correlation coefficients can be found by the formulas:

$$r^{ч}_{YXi} = \frac{r_{YX_i} - r_{X_iX_j} \cdot r_{YX_j}}{\sqrt{1 - r^2_{X_iX_j}} \cdot \sqrt{1 - r^2_{YX_j}}};\ r^{ч}_{X1X2} = \frac{r_{X_1X_2} - r_{X_1Y} \cdot r_{YX_2}}{\sqrt{1 - r^2_{X_1Y}} \cdot \sqrt{1 - r^2_{YX_2}}}$$

For the concerned example, the partial coefficient values equal to:

$r^{ч}_{YX1} = -0.954833$; $r^{ч}_{YX2} = 0.503878$; $r^{ч}_{X1X2} = 0.462724$.

When comparing the matching coefficient and partial coefficients, the following conclusions can be drawn. The volume of counterfeit products has a great impact on the volume of products sold, both in the case of a permanent volume of products produced and in the case of a change in this volume. At a permanent volume of counterfeit products, the strength of the dependence of the volume of products sold on the volume of products produced increases.

For greater confidence in the results obtained, the correlation coefficients can be verified using corresponding statistical hypotheses confirming (or rejecting) their significance.

After conducting a correlation and regression analysis, one can proceed to predict the estimated volume of products that can be sold in December.

Finding the predicted value

All the studies conducted above are necessary to ensure that the prediction made below corresponds to the real situation. The point prediction is determined quite simply, after substituting the possible values of factor features into the regression equation. For the data of the considered example, the possible volume of goods sold in December under the same conditions will be approximately equal to

$$y = 22.3202 - 1.4584 \cdot 10 + 0{,}2899 \cdot 30 = 16.43401.$$

4. Conclusion

In conclusion, it should be noted that before conducting a correlation and regression analysis, it is desirable to check the uniformity of the sample data. If the sample is not homogeneous, then applying correlation and regression methods to its elements can lead to results that are far from reality. In this case, it is recommended to conduct clustering of the sample population [6, 8], i.e. its division

into groups that are homogeneous in a certain sense, and then conduct a correlation and regression analysis of the problem separately for each group.

The considered case is characterized by the availability of the clearly expressed resulting feature, while there are a lot of factor features. The question arises of how to choose the resulting factor from all the factors that have a significant impact. According to the authors, a promising area for further research may be using the method of expert assessments, which will allow identifying the most important features of the development of the situation under study based on the survey of competent experts.

References

Bolshakova, L.V., Litvinenko, A.N., Baturina, E.V., Sidenko, I.K., Ivanov, A.N., Dali, F.A., Shidlovsky, G.L. 2020. Application of the econometric model as a mechanism of management of socio-economic systems. The IIOAB Journal, 11, No S3. pp. 64-71.

Bolshakov, L.V., Smolin, Ya.N., Yakovleva, N. A. 2019. Mathematical-statistical study of the increase in debt to banks of the population. In: Regional Informatics and Information Security. pp. 212-214.

Kasych, A., Vrbka, J., Rowland, Z., Glukhova, V.Modern human resource management models: Values, development approaches, transformation (2020) Quality - Access to Success, 21 (179), pp. 72-79.

Kadochnikova, E.I. 2012. Methodological problems of constructing models of economic growth in the region. Bulletin of Economics, Law, and Sociology., No. 1. pp. 52-56.

Kartavtseva, A.V. 2016. Econometric analysis of inflation in the Russian Federation. Issues of Economics and Management, No. 5(7). pp. 29-33. https://moluch.ru/th/5/archive/44/1513/

Litvinenko, A.N., Bolshakova, L.V., 2020. Methods of applying cluster analysis when performing final qualifying work by students of the St. Petersburg University of the Ministry of Internal Affairs of Russia. Bulletin of the St. Petersburg University of the Ministry of Internal Affairs of Russia, No. 1(85). pp. 208-217.

Primakin, A.I., Bolshakova, L.V. 2012. The method of expert assessments in solving problems of ensuring the economic security of an economic entity. Bulletin of the St. Petersburg University of the Ministry of Internal Affairs of Russia., No. 1. pp. 191-200.

Rubtsov, G.G., Litvinenko, A.N., Bolshakova, L.V., 2020. Trends in the development of domestic innovation policy as exemplified by the Northwestern Federal District. Scientific and Technical Bulletin of the St. Petersburg State Polytechnic University. Economic Sciences, 13(1), pp. 65-78.

Snatenkov, A. A., Timofeeva, T.V. 2017. Statistical study of factors forming overdue debt on loans of the Russian banking sector. Competitiveness in the global world: Economics, science, technologies. No. 7-1. pp. 137-144.

Tarasova, T.A. 2017. Regression analysis of the dependence of GDP growth on the factors of innovative economy. Scientific and methodological electronic journal Concept, No. 12. Retrieved from http://e-koncept.ru/2017/174025htm.

Filimonenko, I.V., 2011. Modeling of the dependence of GDP growth on changes in the employment structure in the Russian economy. Bulletin of the Novosibirsk State University. Series: Socio-Economic Sciences, 11(1), pp. 16-25.