MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY
OF ECONOMICS

STATISTICAL THINKING FOR SCIENCE ABOUT DATA

Guidelines
to laboratory work

for Master's (second) degree
students of speciality 122 "Computer Science"

Kharkiv
S. Kuznets KhNUE
2020

UDC 311(07.034)
S81

Compiled by: O. Rayevnyeva
V. Derykhovska

Затверджено на засіданні кафедри статистики і економічного прогнозування.

Протокол № 10 від 21.04.2020 р.

Самостійне електронне текстове мережеве видання

Statistical Thinking for Science about Data [Electronic resource] :
S81 guidelines to laboratory work tasks for Master's (second) degree students of speciality 122 "Computer Science" / compiled by O. Rayevnyeva, V. Derykhovska. – Kharkiv : S. Kuznets KhNUE, 2020. – 96 p. (English)

Tasks for laboratory work on the academic discipline and guidelines to them are given to help students gain practical skills in the use of the tools of economic and mathematical modeling while studying complex socioeconomic processes and systems.

For Master's (second) degree students of speciality 122 "Computer Science".

UDC 311(07.034)

Introduction

The rapid development and wide application of the newest packages of applied programs and computer technology tools necessitate the formation of a specialist in business intelligence and information systems having new competences aimed at acquiring knowledge and skills in the use of econometric and mathematical modeling for the analysis of complex, mass socioeconomic phenomena and processes in various spheres of activity.

At present, the demand for specialists who combine the competence in the application of intelligent information and computer processing systems, the use of modern software products, IT technologies and technological tools in professional, in particular, entrepreneurial activity with the competences of a business analyst for substantiation and making management and business decisions is the main trend in the national and international labor markets. This discipline is a response to the contemporary needs of the community which provides students with an in-depth understanding of the business context of any socioeconomic processes and enables them to solve the problems associated with analytical work in the IT-industry.

The performance of laboratory work tasks aims to develop students' skills in the extension and deepening of theoretical knowledge and acquisition of professional competences in forecasting of socioeconomic processes and modeling of complex systems.

Studying this discipline enables the student:

to get acquainted with available statistical methods and models;

to identify the main features of modeling and forecasting of complex socioeconomic systems;

to study socioeconomic processes using econometric models, additive prediction models, factor analysis, cluster and discriminant analysis, analysis of weakly formalized situations using expert analysis.

Statistical thinking for science about data is one of the basic disciplines of the master's program "Business Analysis and Information Systems in Entrepreneurship".

Content module 1. The methodological bases of statistical modeling and forecasting

Topic 1. The categorical basis of statistical modeling and forecasting

Laboratory work 1
Introduction to the Statistica 8.0 package. Study of the statistical characteristics of the variational series. Checking the dynamic rows for the law of normal distribution

The purpose of the work is to consolidate the theoretical and practical skills in studying the statistical characteristics of the variation series and to check the law of distribution of the dynamic series in the package Statistica 8.0.

The task is to analyze the variation series using descriptive statistics and verify the law of distribution of the dynamic series.

Guidelines

1. Work should begin with the launch of the Statistica package, which is similar to running other applications – through the START menu or using a shortcut.

The Statistica system window consists of the following main elements: the title bar, the menu bar, the toolbar, the workspace and the status bar.

The title bar contains an icon, the name of the program Statistica, and three buttons for managing the size of the main window: the button minimizing the size of the window; the window resizing button; the button closing the window.

The menu bar occupies the second row of the main module window, and if there is an open data file in the workspace, it contains the so-called drop-down menu: *File*, *Edit*, *View*, *Insert*, *Format*, *Statistics*, *Data mining*, *Graphs*, *Tools*, *Data*, *Window*, *Help*.

The toolbar contains buttons for quick access to the most commonly used menu commands.

The workspace in which the various documents are displayed takes up most of the main window:

1) the spreadsheet with the source data. When you first open Statistica in the workspace, a new 10x10 file with the name *Spreadsheet.sta* opens automatically;

2) the startup window of the statistical analysis module used;

3) electronic spreadsheets with results of analysis;

4) graphing tools;

5) the self-timer window.

The status bar is located at the very bottom of the system window.

Depending on the state in which the system is located, the status bar contains a quick access button to the main statistical modules and menu items, as well as displays different information and allows you to control the operation of the system.

When processing data and constructing graphs, the status bar contains a progress bar that reflects the degree of completeness of data processing and the timer, which reflects the time elapsed since the beginning of the processing.

As an example, consider the value of capital investment in the regions (the source of information is the official site of the State Statistics Service of Ukraine), Table 1.1.

Table 1.1

Capital investment on a regional basis in 2018

No.	Regions of Ukraine	The volume of the capital investment, mln UAH
1	2	3
1	Vinnytsya	17626.5
2	Volyn	8687.0
3	Dnipropetrovsk	60288.6
4	Donetsk	26979.4
5	Zhytomyr	8742.3
6	Zakarpattya	7500.6
7	Zaporizhzhya	15732.0
8	Ivano-Frankivsk	9393.7
9	Kyiv	40713.4
10	Kirovohrad	7181.5
11	Luhansk	3219.3
12	Lviv	28995.5
13	Mikolayiv	10099.2

Table 1.1 (the end)

1	2	3
14	Odesa	23787,8
15	Poltava	18636,7
16	Rivne	7228,0
17	Sumy	7749,9
18	Ternopil	8375,0
19	Kharkiv	23551,3
20	Kherson	8853,2
21	Khmelnytskiy	11274,9
22	Cherkasy	11110,4
23	Chernivtsi	3720,6
24	Chernihiv	8971,3
25	City of Kyiv	200308,3

To do this, in the *File* menu of the Statistica system, select the *New* tab. As a result, the window for creating a new file will open (Fig. 1.1).

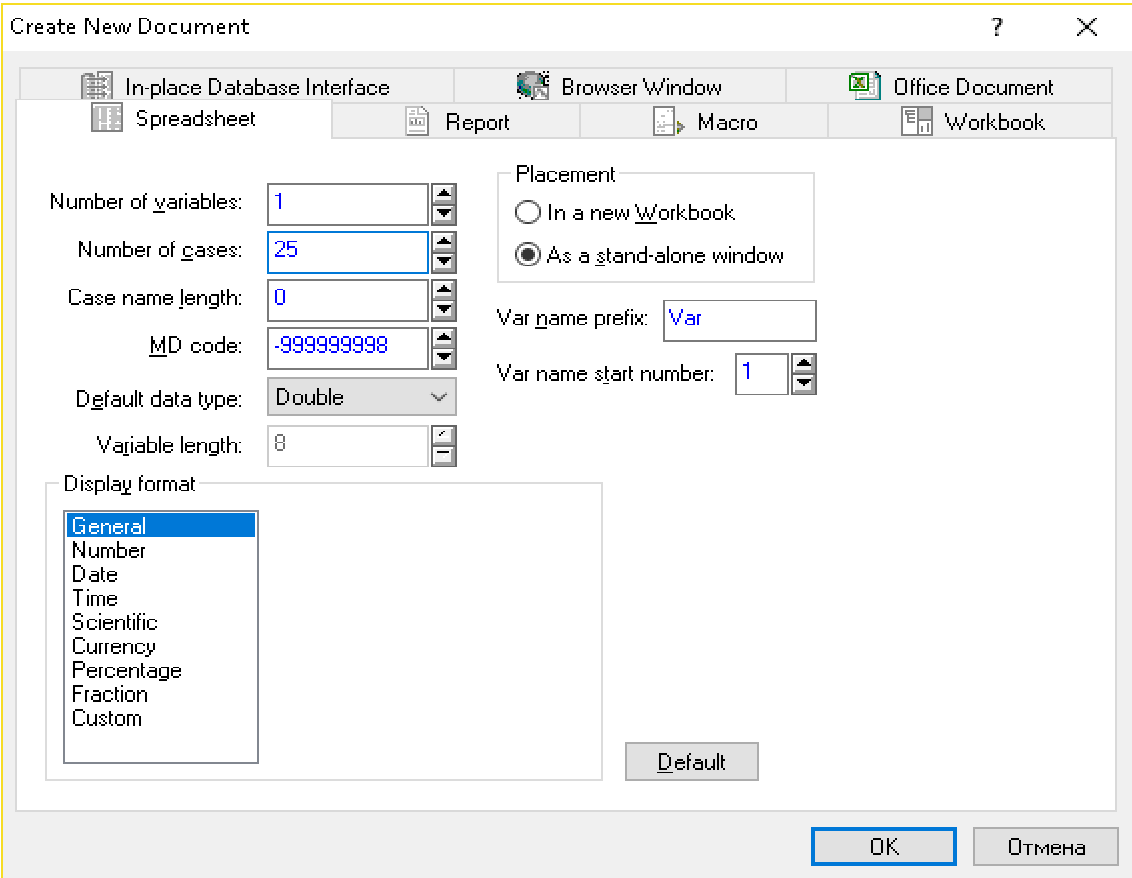


Fig. 1.1. **Creating a new file**

You need to create a spreadsheet with one variable (1st column) and 25 observations (rows) and enter numerical values into it. After that, the spreadsheet will look like this (Fig. 1.2).

Data: Spreadsheet4* (1v by 25c)

	1 The volume of the capital investment
Vinnytsya	17626,5
Volyn	8687
Dnipropetrovsk	60288,6
Donetsk	26979,4
Zhytomyr	8742,3
Zakarpattya	7500,6
Zaporizhzhya	15732
Ivano-Frankivsk	9393,7
Kyiv	40713,4
Kirovohrad	7181,5
Luhansk	3219,3
Lviv	28995,5
Mikolayiv	10099,2
Odesa	23787,8
Poltava	18636,7
Rivne	7228
Sumy	7749,9
Ternopil	8375
Kharkiv	23551,3
Kherson	8853,2
Khmelnytskiy	11274,9
Cherkasy	11110,4
Chernivtsi	3720,6
Chernihiv	8971,3
City of Kyiv	200308,3

Fig. 1.2. **The input data in Statistica 8.0**

In the *File* menu, click *Save* and save the file in the created folder under any name (for example, Regions.sta).

You can change the structure of the spreadsheet using the *Data* menu, choosing a variety of basic monitoring functions and variables:

- add and delete;
- cut and paste;
- sort and standardize;
- transpose and move, etc.

The calculation of the basic numerical characteristics of the investigated variation series can be carried out using descriptive statistics. From the

Statistics menu, select *Basic statistics / Tables*. In the window, select *Descriptive statistics*, which will open a window for calculating the complex descriptive statistics (Fig. 1.3).

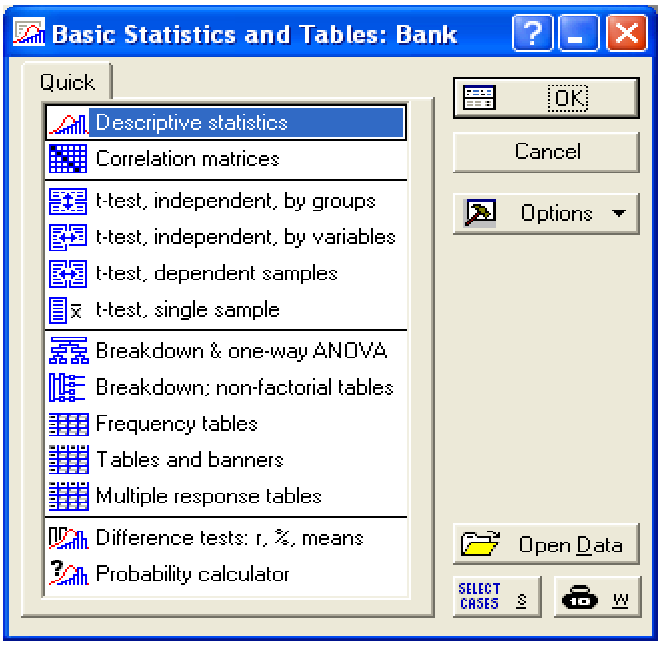


Fig. 1.3. **The window of *Basic statistics***

Next, you need to go to the *Advanced* tab and select the metrics you want to calculate by checking the boxes next to them as shown in Fig. 1.4.

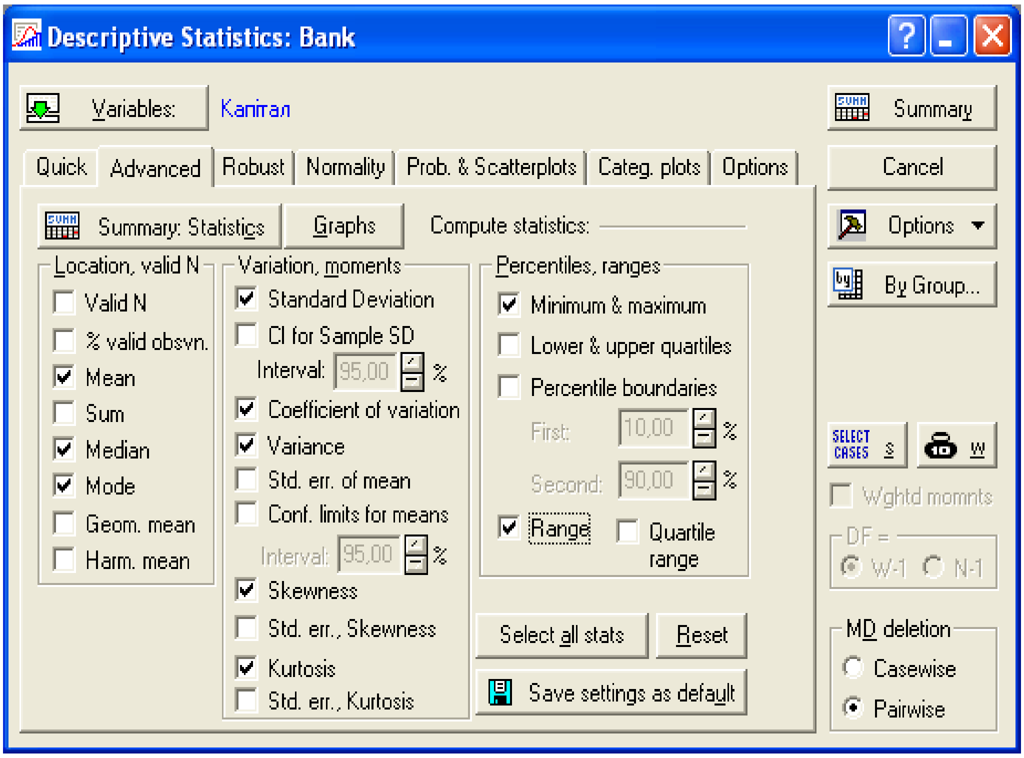


Fig. 1.4. **Choosing the parameters of *Descriptive statistics***

Thus, for the analysis the following indicators are chosen: Mean – arithmetic mean; Median; Mode; Standard deviation; Variance; Skewness; Kurtosis; Minimum & Maximum; Range; Coefficient of variation.

The system will calculate these indicators and present the results in the form of a table (Fig. 1.5) after the button 

is pressed.

	Descriptive Statistics (Spreadsheet4)											
Variable	**Mean**	Median	Mode	Frequency of Mode	Minimum	Maximum	Range	Variance	Std.Dev.	Standard Error	Skewness	Kurtosis
The volume of the capital investment	23149,06	10099,20	Multiple		3219,300	200308,3	197089,0	1,528310E+09	39093,61	7818,722	4,216404	19,22350

Fig. 1.5. **The results of the function *Descriptive statistics***

Statistica has a powerful graphical toolkit: histograms, point and line graphs, two-dimensional and three-dimensional diagrams, etc. Let's consider some types of statistical 2D histograms.

2D Histograms are graphical representations of the frequency distribution of selected variables. For each interval (class) a column is drawn, whose height is proportional to the frequency of the class. The histogram clearly shows which values or ranges of values of the investigated variable are the most frequent ones, how different they are, how the majority of observations are concentrated around the average, whether the distribution is symmetrical or not, whether it has one or several modes. There are several types of histograms:

2D Regular (ordinary) Histograms which are frequency distribution barcharts for a selected variable (in case of more than one variable, a separate diagram is built for each of them);

2D Multiple Histograms which represent the frequency distribution for several variables in a single graph. Frequencies for all variables are placed on the left axis *Y*. The values of all the studied variables are placed on the same axis *X*, which facilitates the comparison of the analyzed variables;

2D Double-Y (double-axis Y) Histograms. The histogram with a double axis *Y* can be considered a combination of two differently scalable histogram components. You can select two different groups of variables for this histogram. For each of the selected variables, the frequency distribution will be depicted, but the frequency of the variables from the first list, called Left *Y* (left axis *Y*), is set on the left axis *Y*, and the frequency variables from the second list, the so-called Right *Y* (right-axis *Y*), are set aside, on the right axis *Y*.

This graph is useful for visual comparison of the distributions of variables with different frequencies;

2D Hanging Bar (hanging columns) Histograms. The hanging column hologram is a "clear criterion for verifying the normality of distribution" which helps to identify the distribution areas where there are differences between observed and expected normal frequencies. It is believed that the columns representing the frequencies observed for the successive value ranges are "suspended" to the most suitable normal curve. If the investigated distribution is close to the normal curve, the lower edges of all columns should form a straight horizontal line.

For example, let's construct a histogram of the distribution of regions depending on the size of capital investments. To do this, in the *Graphs* menu, select *Histograms*, then select the *Capital investment* variable and choose the function *Normal fit*. The result of the analysis is presented in Fig. 1.6.

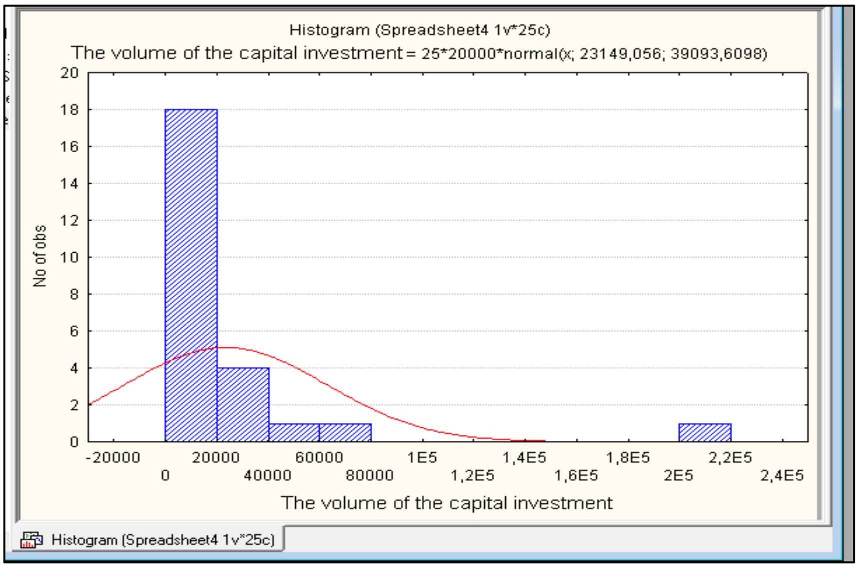


Fig. 1.6. **Visualization of the variable *Capital investment***

2. In order to verify the normal distribution law, it is necessary to investigate the index of industrial production in the period from 2017 to 2018 (Table 1.2).

2.1. Calculation of basic statistics.

In order to perform computational procedures, you must enter the *Statistics / Basic Statistics / Tables* menu. In the window, you need to select *Descriptive Statistics*. In the startup panel of the module, using the *Variable* button (variables), you must specify the output parameters of the model. In this case, it is the *Index* column (Fig. 1.7).

Table 1.2

The index of industrial production in the period from 2017 to 2018

Period (months of 2016 – 2017)	Indices of industrial production in Ukraine	Period (months of 2016 – 2017)	Indices of industrial production in Ukraine
01.16	99.0	01.17	105.6
02.16	108.6	02.17	95.4
03.16	105.8	03.17	100.4
04.16	104.5	04.17	97.3
05.16	104.5	05.17	99.3
06.16	104.5	06.17	93.9
07.16	101.0	07.17	98.0
08.16	103.8	08.17	101.2
09.16	97.1	09.17	98.7
10.16	102.6	10.17	103.8
11.16	100.5	11.17	99.6
12.16	102.3	12.17	97.4

At the next step, select the basic statistics for calculations, as shown in Fig. 1.7, such as: *Valid N* (number of observations), *Mean*, *Sum*, *Median*, *Mode*, *Standard Deviation*, *Variance*, *Std. err of mean* (mean error), *Skewness* (asymmetry coefficient), *Std. err of skewness* (error of asymmetry coefficient), *Kurtosis* (coefficient of excess), *Std. err of Kurtosis*, *Minimum & Maximum*, *Range* (Sample Swap).

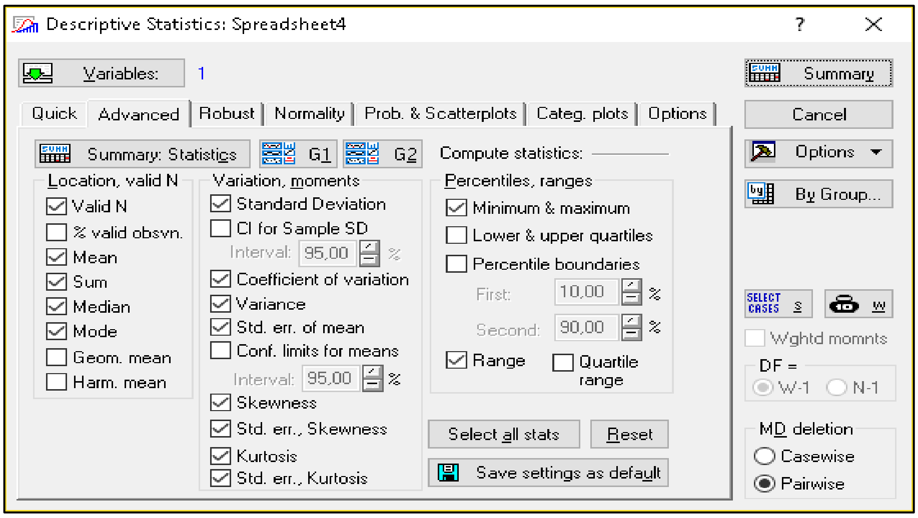


Fig. 1.7. **Choosing the parameters of the *Descriptive statistics***

The results of the calculation are shown in Fig. 1.8.

Variable	Descriptive Statistics (Spreadsheet4) Valid N	Mean	Median	Mode	Frequency of Mode	Sum	Minimum	Maximum	Range	Variance	Std.Dev.	Coef.Var.	Standard Error	Skewness	Std.Err. Skewness	Kurtosis	Std.Err. Kurtosis
Indices of industrial production in Ukraine	24	101,0333	100,7500	104,5000	3	2424,800	93,90000	108,6000	14,70000	13,30580	3,647711	3,610404	0,744586	0,048676	0,472261	-0,520857	0,917777

Fig. 1.8. **The results of the function *Descriptive statistics***

Based on the calculations made it can be concluded that the average index of industrial production is 101.03 %. In the period from January 2017 to December 2018, the highest value of the index of industrial output was 108.6 %, and the smallest one was 93.9 %. The median value is closer to the mean value, indicating that the spread of values is not large. The definition of the distribution form is carried out using the following criteria:

• the value of the asymmetry ratio exceeds 0 in the case of the right asymmetry and less than 0 in the case of the left asymmetry. In our case, a right-side asymmetry is observed;

• the value of the coefficient of excess is greater than 0 in the case of a spin-off distribution and less than 0 in the case of a flat-distributive distribution. The sample being studied is flat-top, because the coefficient of excess is negative.

2.2. Construction of a distribution polygon (frequency polygon).

To do this, go to the *Graphs / 2D Graphs / Line Plots* menu (variables) and select a variable.

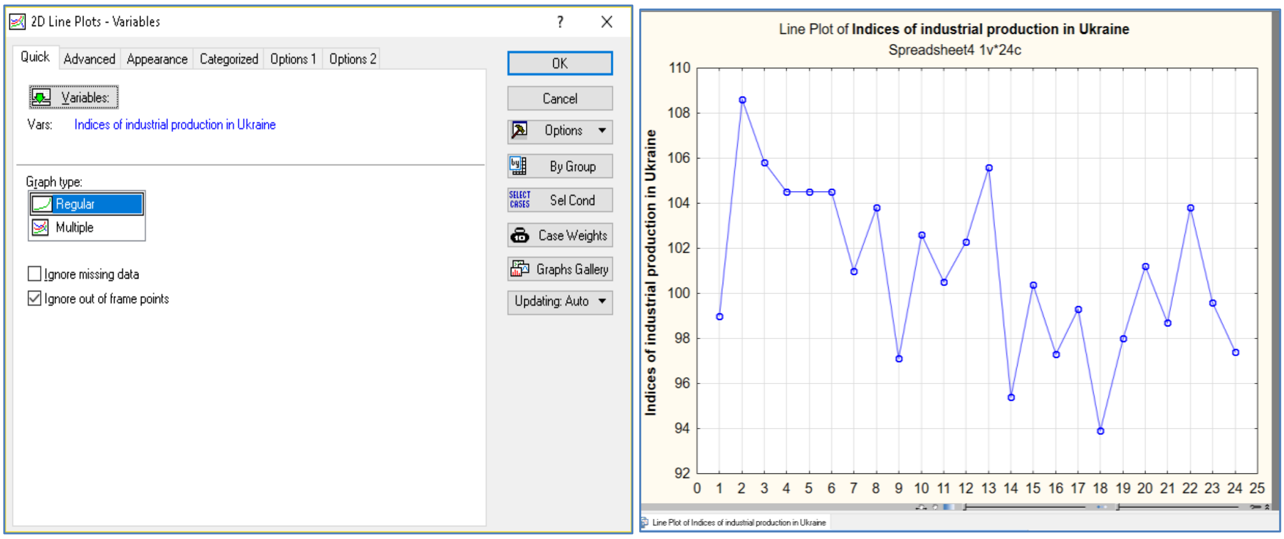


Fig. 1.9. **The distribution polygon**

From Fig. 1.9 it is clear that for the indicator under research it is quite difficult to characterize a certain tendency. However, the highest growth of the index of industrial output was noted in February 2017, and a significant reduction – in June 2018.

3. Check the sample for the normal distribution law.

3.1. Calculation of the number of intervals of grouping. The number of intervals of grouping is calculated according to the following formula:

$$m = 1 + 3.322\lg n, \tag{1.1}$$

where n is the length of the dynamic series.

$$m = 1 + 3.322\lg 24 = 5.585 \approx 6.$$

3.2. Further analysis is carried out as part of the sample verification for the normal distribution law. Calculate the Pearson criterion to draw a final conclusion on the nature of the distribution of magnitude. To do this, in the main menu *Statistics* select the *Distribution Fitting* module. Fig. 1.10 shows the choice of the distribution law that is being tested.

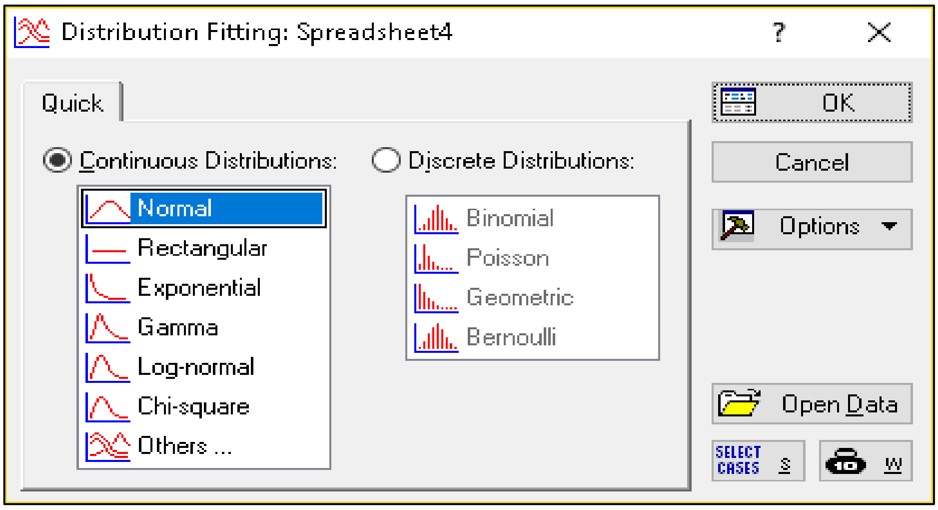


Fig. 1.10. **Choosing the parameters in the *Distribution Fitting* module**

Next, define the parameters of the calculation, as shown in Fig. 1.11. The parameter window indicates the number of intervals, the average index of industry, the minimum and maximum value of the indicator.

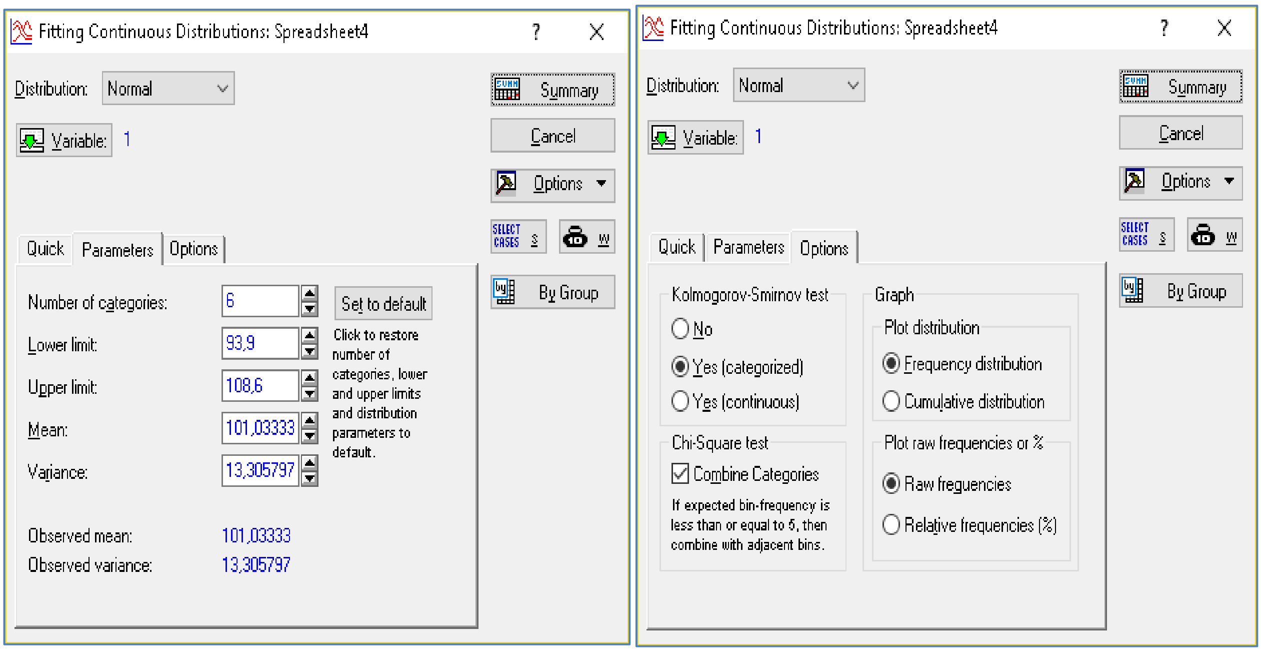


Fig. 1.11. **Choosing the parameters in the *Distribution Fitting* module**

The results of the calculations are shown in Fig. 1.12.

Variable: **Indices of industrial production in Ukraine**, Distribution: Normal (Spreadsheet4)
Kolmogorov-Smirnov d = 0,10096,
Chi-Square = 3,76953, df = 1 (adjusted) , p = 0,05219

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 96,35000	2	2	8,33333	8,3333	2,390081	2,39008	9,95867	9,9587	-0,39008
98,80000	5	7	20,83333	29,1667	4,094339	6,48442	17,05975	27,0184	0,90566
101,25000	7	14	29,16667	58,3333	6,083958	12,56838	25,34983	52,3682	0,91604
103,70000	2	16	8,33333	66,6667	5,854653	18,42303	24,39439	76,7626	-3,85465
106,15000	7	23	29,16667	95,8333	3,648514	22,07155	15,20214	91,9648	3,35149
< Infinity	1	24	4,16667	100,0000	1,928454	24,00000	8,03523	100,0000	-0,92845

Fig. 1.12. **The results of the calculation**

In the normal distribution, the values of the Pearson criterion and the Kolmogorov – Smirnov criteria are less than the critical value. For the degree of freedom k - p - 1 = 4 (where k is the number of distribution intervals, p is the number of parameters to be checked) and the significance level α = 0.01, the table value of the Pearson criterion is 13.3. Calculations show that Pearson's criterion is 3.77 for the industrial production index. The critical value of the Kolmogorov – Smirnov criterion with the length of the sample of 24 observation and α = 0.01 is 0.3255, and the estimated index of industrial output is 0.101. Thus, we can conclude that the index of industrial output is distributed according to the normal law.

For clarity, we present the results graphically (Fig. 1.13).

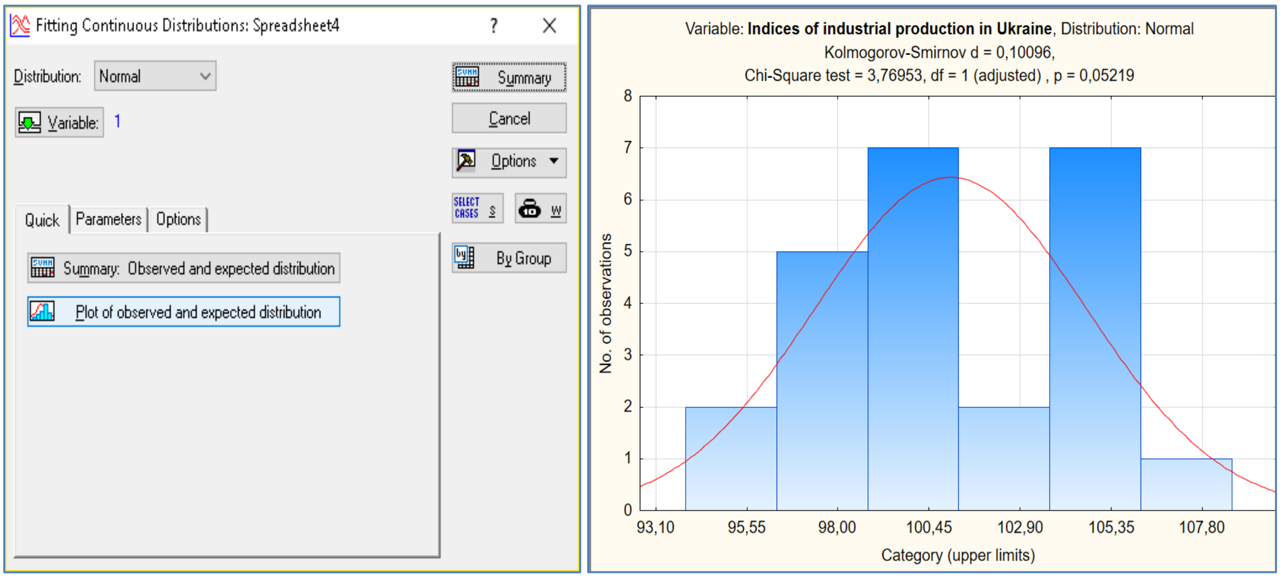


Fig. 1.13. **The distribution histogram with Pearson and Kolmogorov – Smirnov criteria**

Also, graphic confirmation of distribution normality can be verified by initializing the normal probability plot (Fig. 1.14) and using the *Normality* tab (Fig. 1.15).

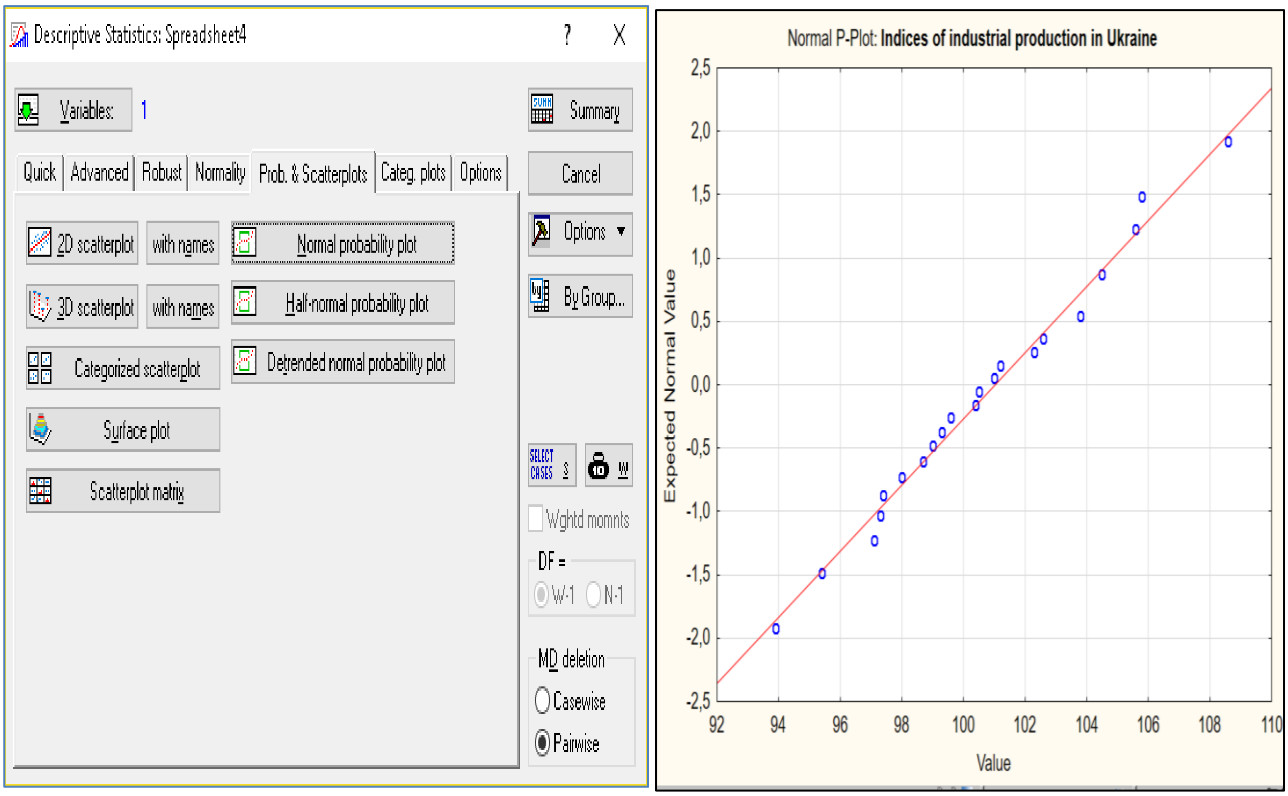


Fig. 1.14. **The normal probability plot**

According to Fig. 1.14 the actual values of the investigated index of location is closely along the theoretical normal line, hence the hypothesis of normality does not deviate, which indicates the normal law of data distribution.

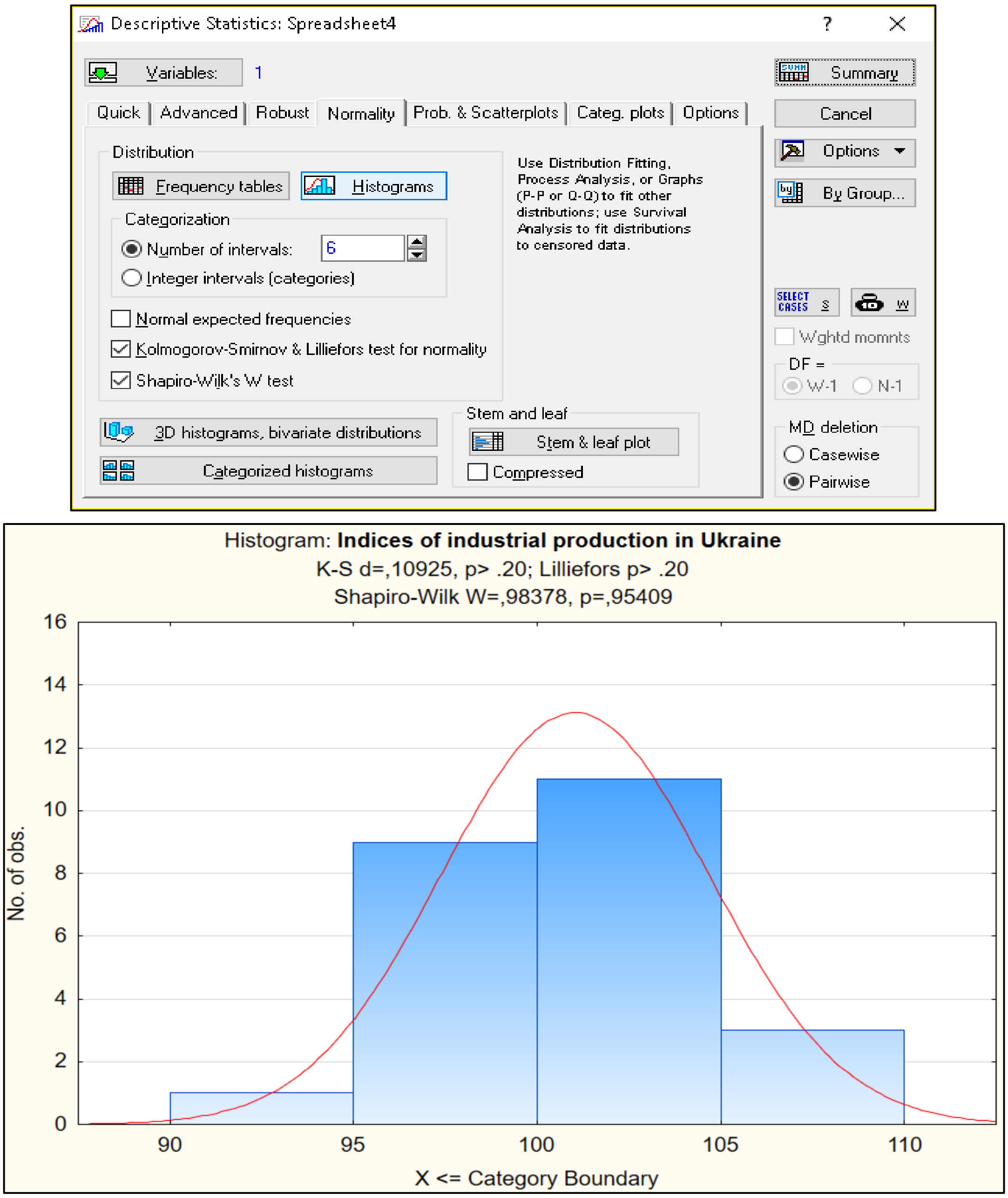


Fig. 1.15. **The histogram with the normal distribution curve**

Tasks for independent work

It is necessary to find spatial one-dimensional data (at least 30 observations) and perform statistical analysis using descriptive statistics and graphic procedures, check your ranks for normal distribution law and draw conclusions (give the economic interpretation of the results).

Topic 2. Regression models as a means of researching economic processes

Laboratory work 2
Development of a single-factor and multiple regression model

The purpose of the work is to consolidate the theoretical and practical material, to acquire skills in the development and analysis of simple and complex econometric models in the package Statistica 8.0.

The task is to check the existence of a linear relationship between the macroeconomic indicators in the Multiple Regression Statistica module.

Guidelines

Task 1. Development of a single-factor regression model.

It is necessary: 1) to develop a linear econometric model and determine all its characteristics (model parameters, mean square deviation of model parameters, variance and mean square deviation of model errors, coefficients of correlation and determination); 2) to check the statistical significance of the model parameters and correlation coefficient according to Student's criterion; to check the adequacy of the model according to Fisher's criterion; 3) to calculate the theoretical values of the dependent variable and model error; to plot a linear function graph with confidence intervals; to plot a histogram and error distribution graph on the normal probability paper; 4) to calculate the predictive value of the dependent variable and confidence intervals of the change if the value of the independent indicator is known; 5) to draw conclusions about the adequacy of the built model, give an economic interpretation of this dependence and the possibility of using the model.

The input data are given in Table 2.1.

1. In the package Statistica 8.0, we introduce the initial data – the level of the GDP (y) and the direct foreign investment in the economy of Ukraine (x). In addition, the factor x is a factor sign, since it affects the resultant (the effect sign) which is y (Fig. 2.1).

Table 2.1

The level of the GDP and direct investment in the economy of Ukraine in 2001 – 2018

Years	GDP, billion UAH	Direct investments, million dollars
2001	148.9	423.6
2002	159.8	483.5
2003	162.5	896.9
2004	165.8	1438.2
2005	186.5	2063.6
2006	192.5	2810.7
2007	198.9	3281.8
2008	221.6	3875
2009	225.8	4555.3
2010	267.3	5471.8
2011	345.1	6794.4
2012	441.5	9047
2013	544.2	16890
2014	720.7	21607.3
2015	948.1	29542.7
2016	913.3	35616.4
2017	1082.6	40053
2018	1316.6	44806

Data: Spreadsheet2* (2v by 18c)

	1 GDP (y)	2 Direct investments (x)
1	148,9	423,6
2	159,8	483,5
3	162,5	896,9
4	165,8	1438,2
5	186,5	2063,6
6	192,5	2810,7
7	198,9	3281,8
8	221,6	3875
9	225,8	4555,3
10	267,3	5471,8
11	345,1	6794,4
12	441,5	9047
13	544,2	16890
14	720,7	21607,3
15	948,1	29542,7
16	913,3	35616,4
17	1082,6	40053
18	1316,6	44806

Fig. 2.1. **The initial data**

2. First let's analyze the influence of the factor sign on the resultant. For this purpose, in the Statistica 8.0 package, we select *Basic statistics* and *Correlation matrices*. Fig. 2.2 shows the main steps of building a correlation matrix.

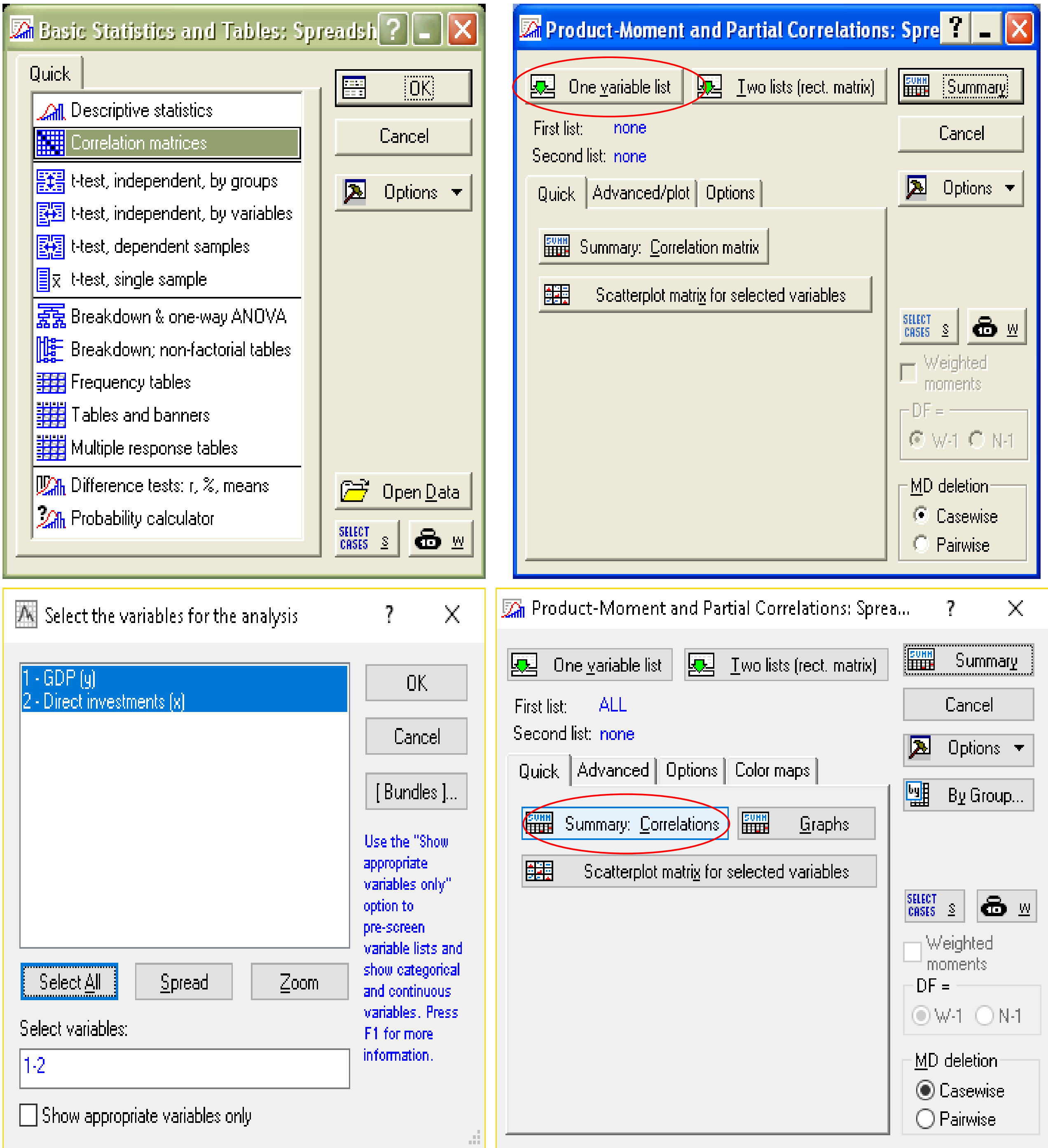


Fig. 2.2. **The steps of building the *correlation matrix***

Then press *Summary: Correlations* and get the results (Fig. 2.3).

Variable	Correlations (Spreadsheet2) Marked correlations are significant at p < ,05000 N=18 (Casewise deletion of missing data) Means	Std.Dev.	GDP (y)	Direct investments (x)
GDP (y)	457,87	374,08	1,000000	0,992381
Direct investments (x)	12758,73	14941,19	0,992381	1,000000

Fig. 2.3. **The correlation matrix**

The results of building the correlation matrix indicate that the value of the correlation coefficient is 0.992, which suggests a high level of dependence between the GDP and direct investment.

3. Next, we proceed to building a one-factor linear regression. To start computational procedures, you must enter the *Statistics / Multiple Regression* menu. After confirming the module selection, a module launcher appears, where you need to set the variables for analysis. In the window that appears, we select the *Dependent* and *Independent Variables* in order to build a simple one-factor model. In Fig. 2.4, the stages of building a one-factor model are presented.

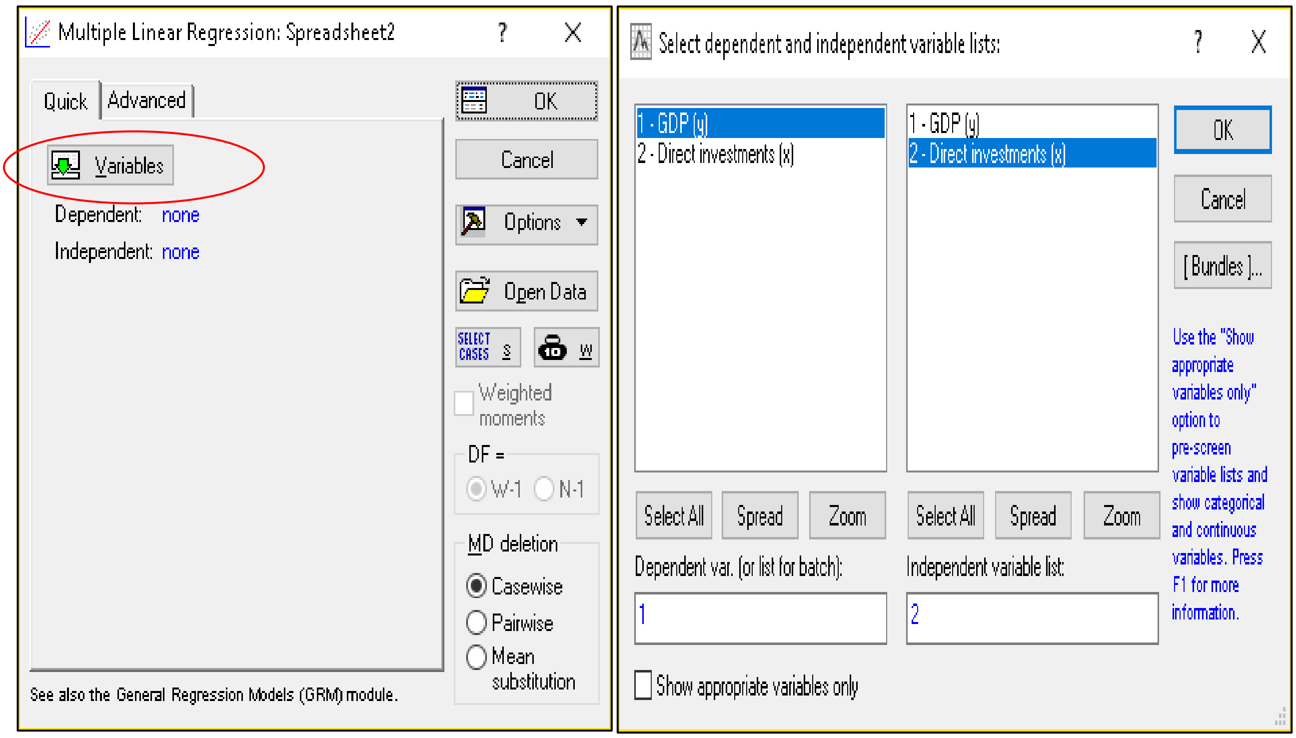


Fig. 2.4. **The stages of building a one-factor model**

When the OK button is pressed, a dialog window (Fig. 2.5) appears with the results of the linear econometric model. The upper part of the window contains basic information about the model, at the bottom there are functional buttons that allow you to comprehensively consider the results of the analysis.

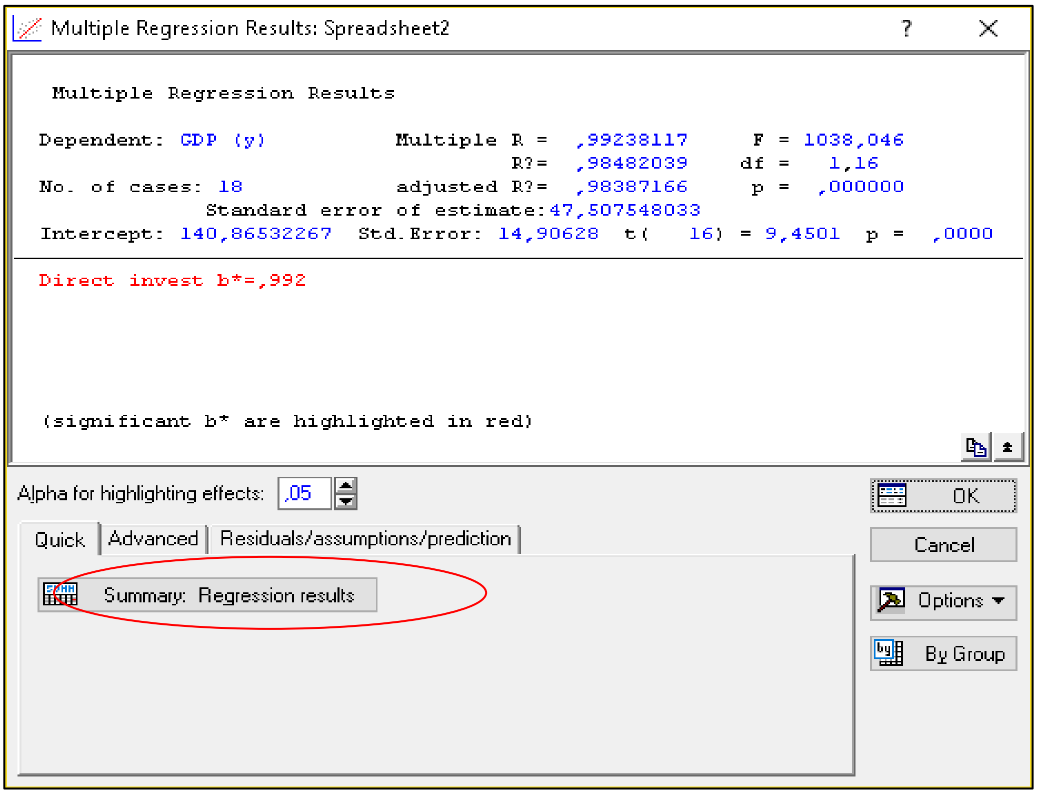


Fig. 2.5. **The results of the linear econometric model**

The characteristics of the model and the degree of their adequacy can be obtained by clicking the *Summary: Regression results* button. The results of building a one-factor econometric model are shown in Fig. 2.6.

	Regression Summary for Dependent Variable: GDP (y) (Sprea R= ,99238117 R?= ,98482039 Adjusted R?= ,98387166 F(1,16)=1038,0 p<,00000 Std.Error of estimate: 47,508					
N=18	b*	Std.Err. of b*	b	Std.Err. of b	t(16)	p-value
Intercept			140,8653	14,90628	9,45006	0,000000
Direct investments (x)	0,992381	0,030801	0,0248	0,00077	32,21871	0,000000

Fig. 2.6. **The regression results**

The obtained results indicate the following:

• the coefficient of multiple correlation (R) is 0.9923. The measured coefficient has the limits from -1 to +1 (if the size of R is close to 1, then the obtained model is adequate and can be used for the analysis and prediction of economic processes);

• the model's determination coefficient (R^2) is 0.9848 (if the size of R^2 is close to 1, then the obtained model is adequate and can be used for the analysis and prediction of economic processes);

• the adjusted determination coefficient based on the number of observations and the number of parameters is 0.9838 (adjusted R^2);

• the Fisher's eligibility criterion F (1, 16) = 1038 (the Fisher's criterion is used to assess the statistical significance of the coefficient of determination. If the values obtained are more than the tabular ones, then R^2 is significant and the model is adequate);

• b (a_1, a_2) = (140.87; 0.0248) are the parameters of the model;

• the mean square deviation of the model parameters is (14.09; 0.0248);

• t (28) = (9.45; 32.22) the significance of the parameters according to the Student's criterion (the Student's criterion is used to assess the statistical significance of the coefficient of correlation. If the values obtained are more than the tabular ones, then R is significant and the model is adequate).

The analysis of the above results shows that the model is adequate and has the following general view:

$$Y = 140.87 + 0.0248\,X\,.$$

Let's plot a linear function graph with confidence intervals. To do this, you must specify the variables, the level line and the confidence intervals in the *Graphs / Scatterplots* menu (Fig. 2.7):

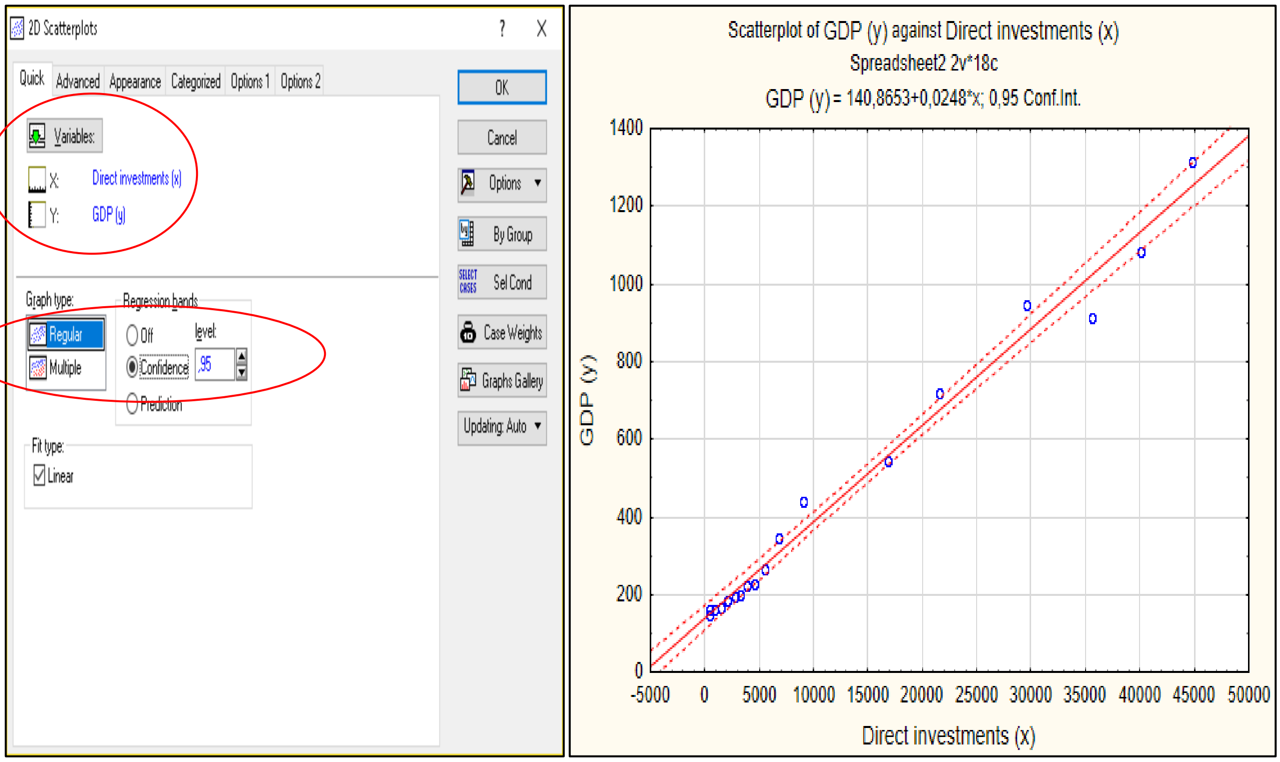


Fig. 2.7. **The linear function graph with the confidence intervals**

The analysis of the graph proves high quality of the built model and the correspondence of the model values to the actual ones.

To calculate and analyze the model's residuals, at the bottom of the window of the results of the regression model there is the option *Perform residual analysis* (Fig. 2.8).

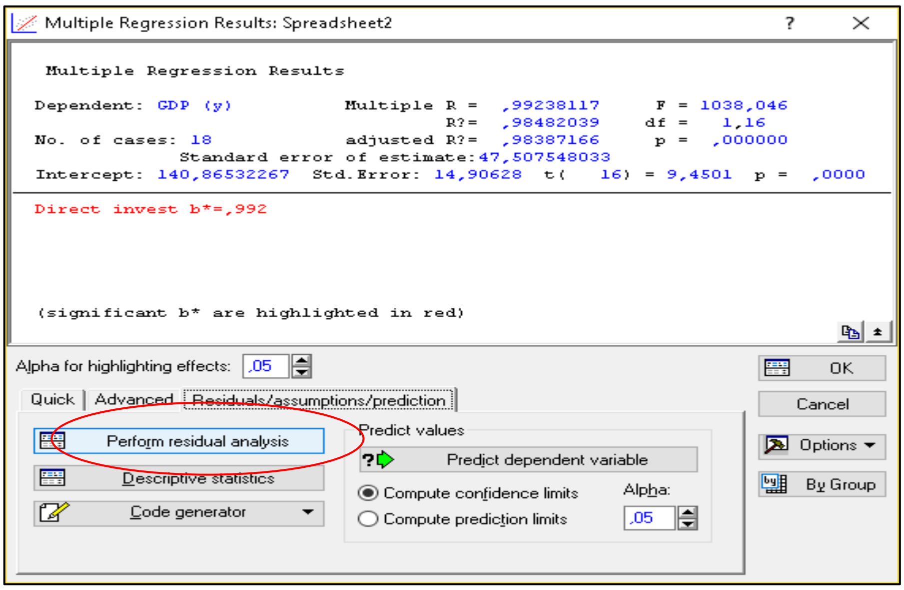


Fig. 2.8. **Choosing the option *Perform residual analysis***

Using this option, we get a menu to analyze the model errors (Fig. 2.9).

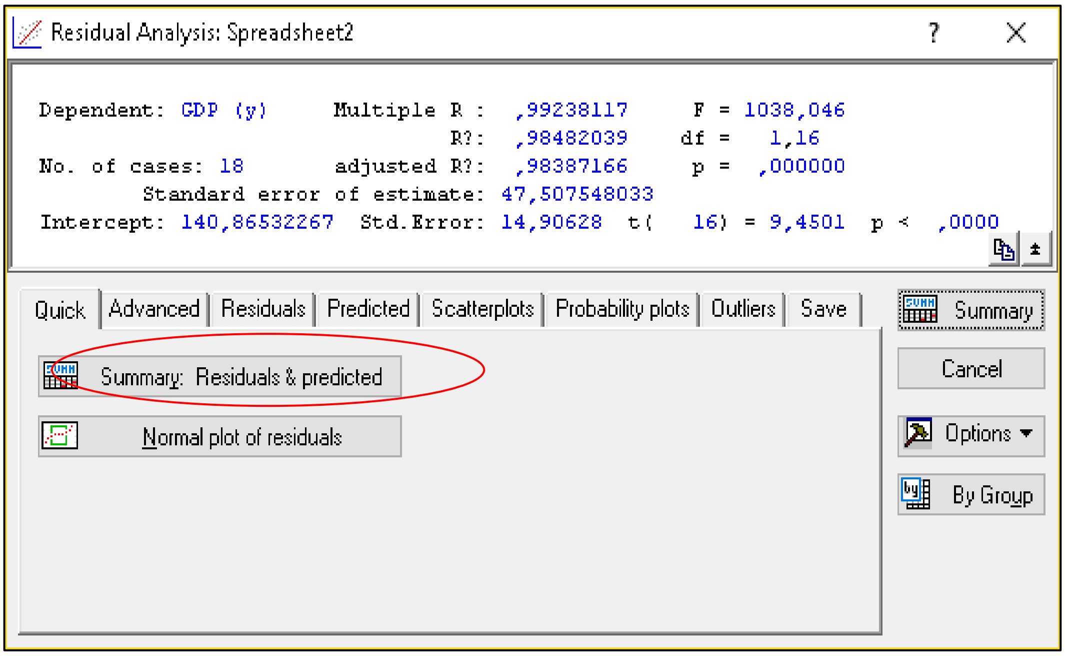


Fig. 2.9. **The mode *Residual analysis***

The Summary Analysis button: *Residuals and Predicted* allows you to get a table containing the actual values of the dependent variable (*Observed value*), its theoretical values (*Predicted value*) and model errors (*Residual*) (Fig. 2.10).

Predicted & Residual Values (Spreadsheet2)
Dependent variable: GDP (y)

Case No.	**Observed Value**	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	148,900	151,390	-2,490	-0,825579	-0,05242	14,69271	0,681581	-2,754	0,000161
2	159,800	152,879	6,922	-0,821570	0,14569	14,66284	0,674977	7,650	0,001235
3	162,500	163,150	-0,650	-0,793902	-0,01368	14,45907	0,630280	-0,716	0,000011
4	165,800	176,599	-10,799	-0,757673	-0,22732	14,19866	0,574068	-11,858	0,002783
5	186,500	192,138	-5,638	-0,715815	-0,11868	13,90732	0,512392	-6,167	0,000722
6	192,500	210,701	-18,201	-0,665813	-0,38311	13,57356	0,443306	-19,819	0,007103
7	198,900	222,406	-23,506	-0,634282	-0,49478	13,37159	0,402314	-25,528	0,011437
8	221,600	237,145	-15,545	-0,594580	-0,32720	13,12715	0,353525	-16,830	0,004791
9	225,800	254,048	-28,248	-0,549048	-0,59459	12,86114	0,301454	-30,481	0,015085
10	267,300	276,819	-9,519	-0,487708	-0,20037	12,52860	0,237859	-10,231	0,001613
11	345,100	309,681	35,419	-0,399187	0,74555	12,10549	0,159351	37,879	0,020638
12	441,500	365,649	75,850	-0,248423	1,59660	11,55770	0,061714	80,622	0,085226
13	544,200	560,519	-16,319	0,276502	-0,34350	11,64205	0,076453	-17,361	0,004010
14	720,700	677,726	42,974	0,592226	0,90457	13,11302	0,350732	46,518	0,036523
15	948,100	874,891	73,209	1,123335	1,54099	17,11485	1,261882	84,127	0,203488
16	913,300	1025,800	-112,500	1,529842	-2,36804	20,88318	2,340418	-139,444	0,832367
17	1082,600	1136,033	-53,433	1,826780	-1,12472	23,84183	3,337125	-71,421	0,284608
18	1316,600	1254,127	62,473	2,144894	1,31501	27,13248	4,600569	92,714	0,621139
Minimum	148,900	151,390	-112,500	-0,825579	-2,36804	11,55770	0,061714	-139,444	0,000011
Maximum	1316,600	1254,127	75,850	2,144894	1,59660	27,13248	4,600569	92,714	0,832367
Mean	457,872	457,872	-0,000	0,000000	-0,00000	15,26518	0,944444	-0,172	0,118497
Median	246,550	265,433	-7,579	-0,518378	-0,15952	13,74044	0,477849	-8,199	0,009270

Fig. 2.10. **The summary analysis: the residuals and the predicted values**

Since the basic hypothesis concerning a random variable says that the errors should be distributed according to the normal distribution law, let's present the graph of the model errors on the normal probability paper (*Residuals / Normal plot of residuals*) (Fig. 2.11).

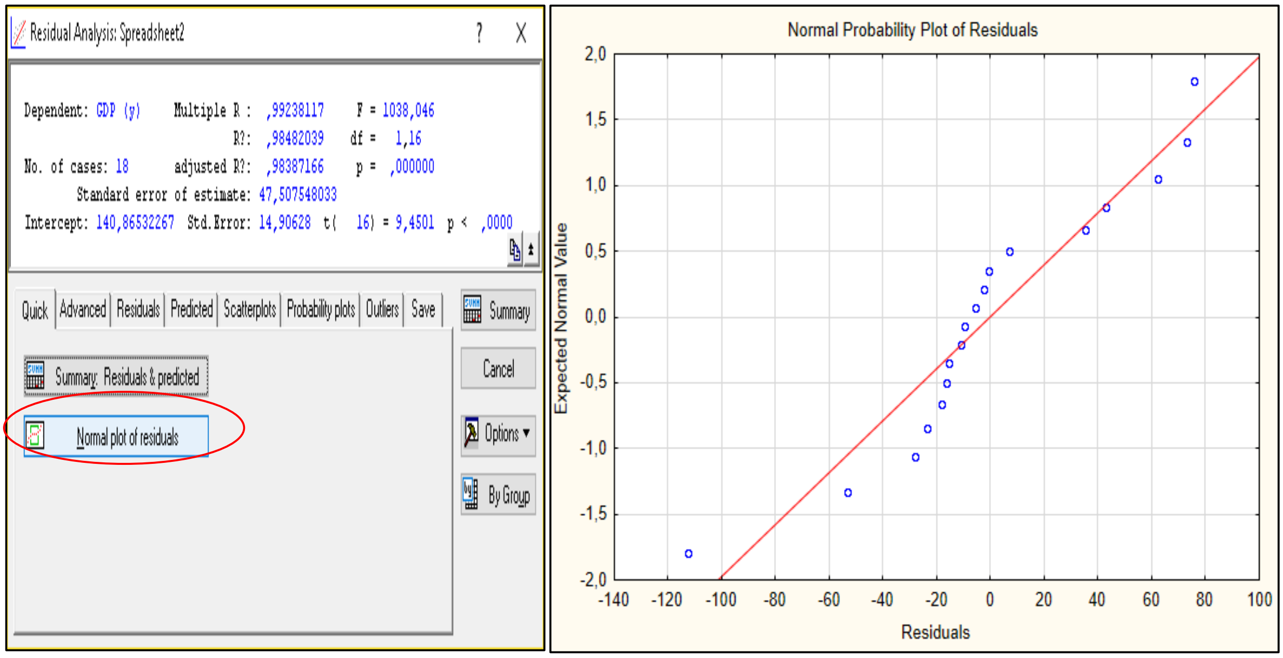


Fig. 2.11. **The normal plot of residuals**

According to the graph, it is difficult to conclude on the law of the distribution of model errors, so for a more detailed analysis we will plot a histogram of their values with plotting a graph of the normal distribution law (Fig. 2.12).

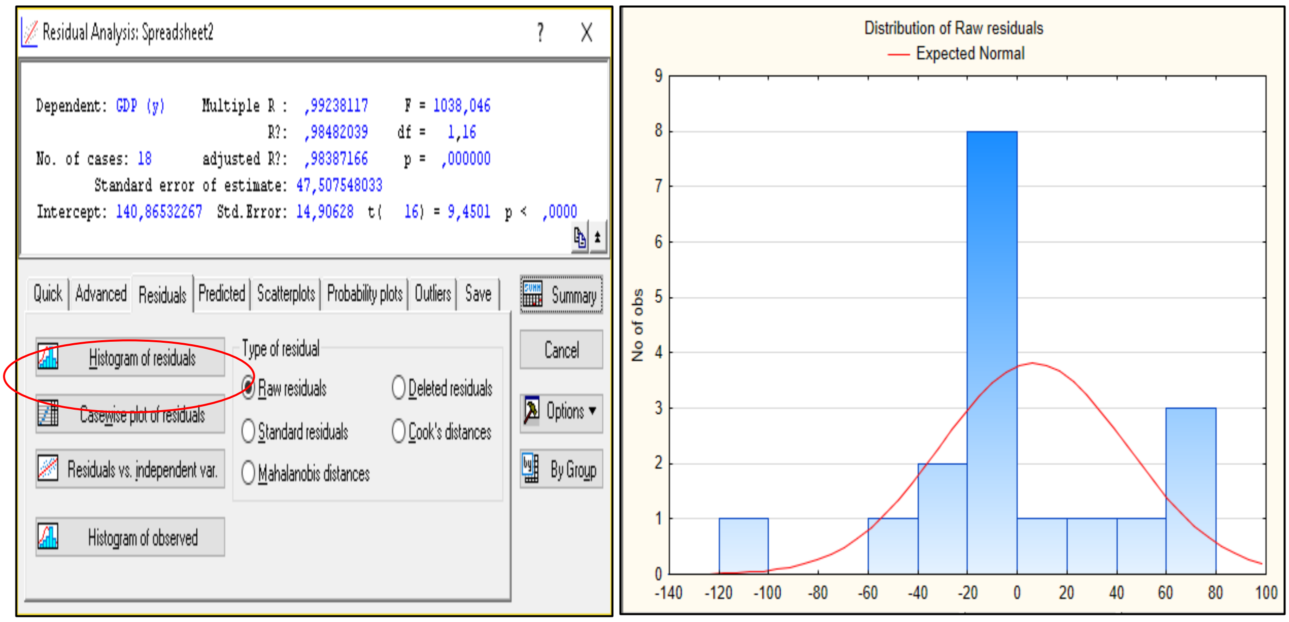


Fig. 2.12. **The histogram of the distribution of the model errors**

4. Since the model is adequate and its parameters are significant, the model can be used to make a forecast. To calculate the predictive values of a dependent variable, the *Predict dependent variable* option is at the bottom of the regression analysis results window (Fig. 2.13).

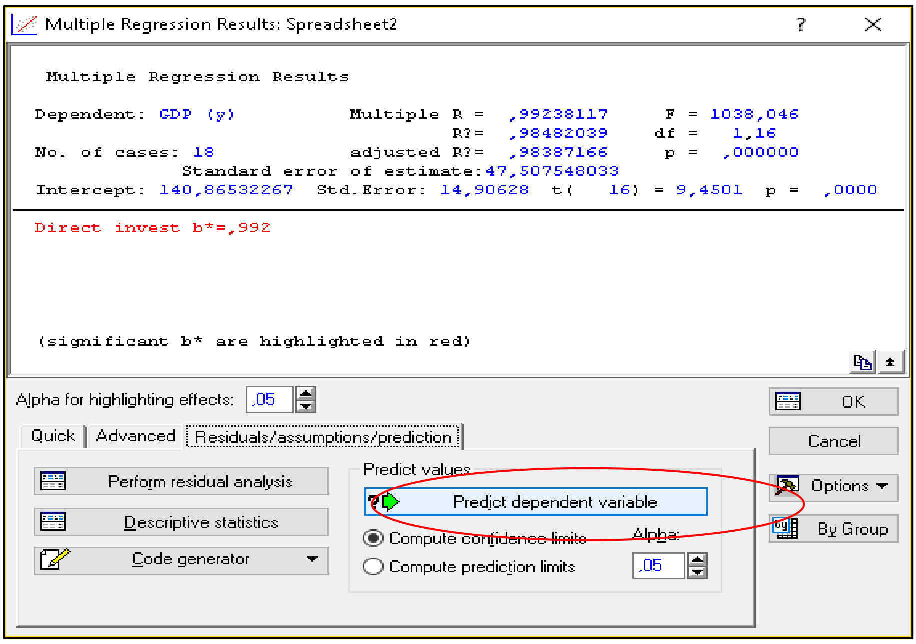


Fig. 2.13. **Choosing the option *Predict dependent variable***

Initiating the appropriate option, you need to specify the amount of direct investment by 2019 (Fig. 2.14).

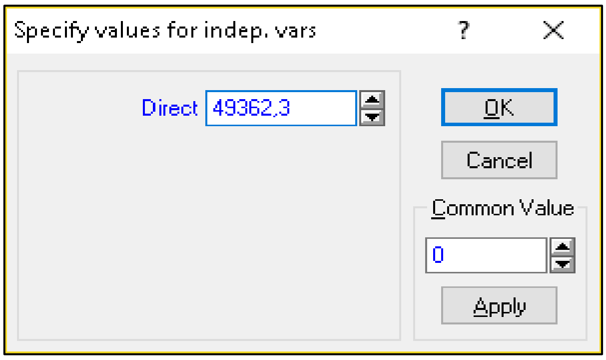


Fig. 2.14. **The input parameter**

The results of the forecasting are presented in the form of a table, which additionally presents the model parameters and the value of the confidence interval of the forecast (Fig. 2.15).

Variable	Predicting Values for (Spreadsheet2) variable: GDP (y)		
	b-Weight	Value	b-Weight * Value
Direct investments (x)	0,024846	49362,30	1226,469
Intercept			140,865
Predicted			1367,334
-95,0%CL			1302,958
+95,0%CL			1431,711

Fig. 2.15. **The results of the forecast**

The forecast value of the GDP *(Predicted)* is 1367.334; the confidence interval of the predicted values is 1302.958 < u <1431.711.

*Conclusion***.** In the work, the analysis of a one-factor linear econometric model of significant dependence of GDP on direct foreign investment in Ukraine was carried out. The developed model is adequate and can be used to make a forecast and evaluate the impact of the exogenous variable.

Task 2. Building a multiple regression model.

To check the existence of linear multiplicity of the link between GDP and socioeconomic indicators (Table 2.2 shows the values for Ukraine in 2005 – 2018), it is necessary: 1) to build a linear multifactor econometric model of the influence of socioeconomic indicators on GDP and determine all its characteristics (model parameters, mean square deviation of model parameters, variance and mean square deviation of model errors, coefficients of multiple correlation and determination); 2) to check the statistical significance of the model parameters, the coefficient of multiple correlation; check the adequacy of the model according to the Fisher criterion; 3) to calculate the theoretical values of the dependent variable and the model error; to plot a linear function graph with confidence intervals; 4) to calculate the predictive value of the dependent variable and confidence intervals of the change if the value of the independent indicators is known; 5) to draw conclusions about the adequacy of the built multifactorial model, and give an economic interpretation of the model as a whole.

Table 2.2

The input data for building a multifactor econometric model

Years	Production output, thou UAH (X1)	Volume of retail turnover of enterprises (legal entities), mln (X2)	Average cost per month per household, UAH (X3)	Direct investments, million dollars (X4)	The GDP, billion UAH (У)
2005	226 358	19 317	395.6	2063.6	186.5
2006	356 842	22 151	426.5	2810.7	192.5
2007	373 893	28 757	541.3	3281.8	198.9
2008	460 520	34 417	607	3875	221.6
2009	504 008	39 691	658.3	4555.3	225.8
2010	603 704	49 994	736.8	5471.8	267.3
2011	809 988	67 556	903.5	6794.4	345.1
2012	995 630	94 332	1229.4	9047	441.5
2013	1 182 179	129 952	1442.8	16890	544.2
2014	1 565 055	178 233	1722	21607.3	720.7
2015	2 072 172	246 903	2590.4	29542.7	948.1
2016	1 955 685	230 955	2754.1	35616.4	913.3
2017	2 388 289	280 890	3072.7	40053	1082.6
2018	2 496 365	350 059	3456	44806	1316.6

1. According to the algorithm of building a one-factor regression model, calculations were made for a multifactor model (Fig. 2.16).

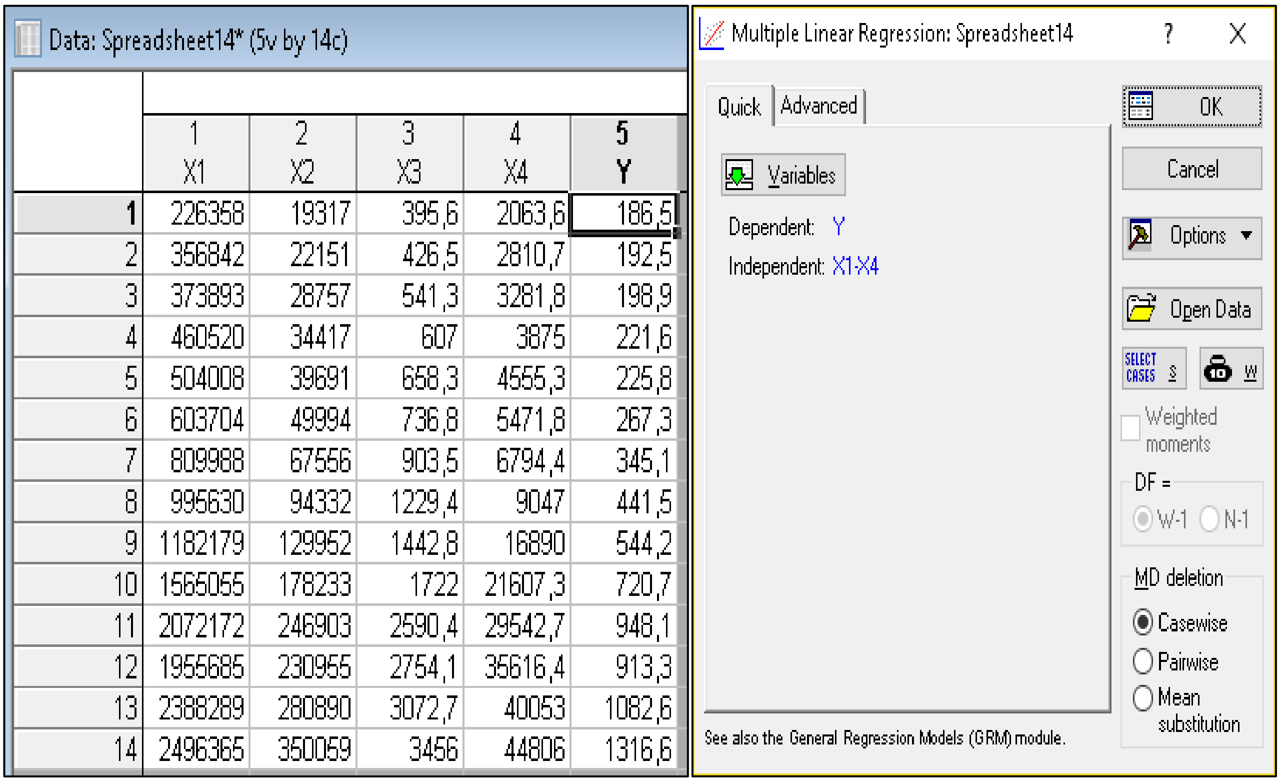


Fig. 2.16. **The stages of building a multifactor model**

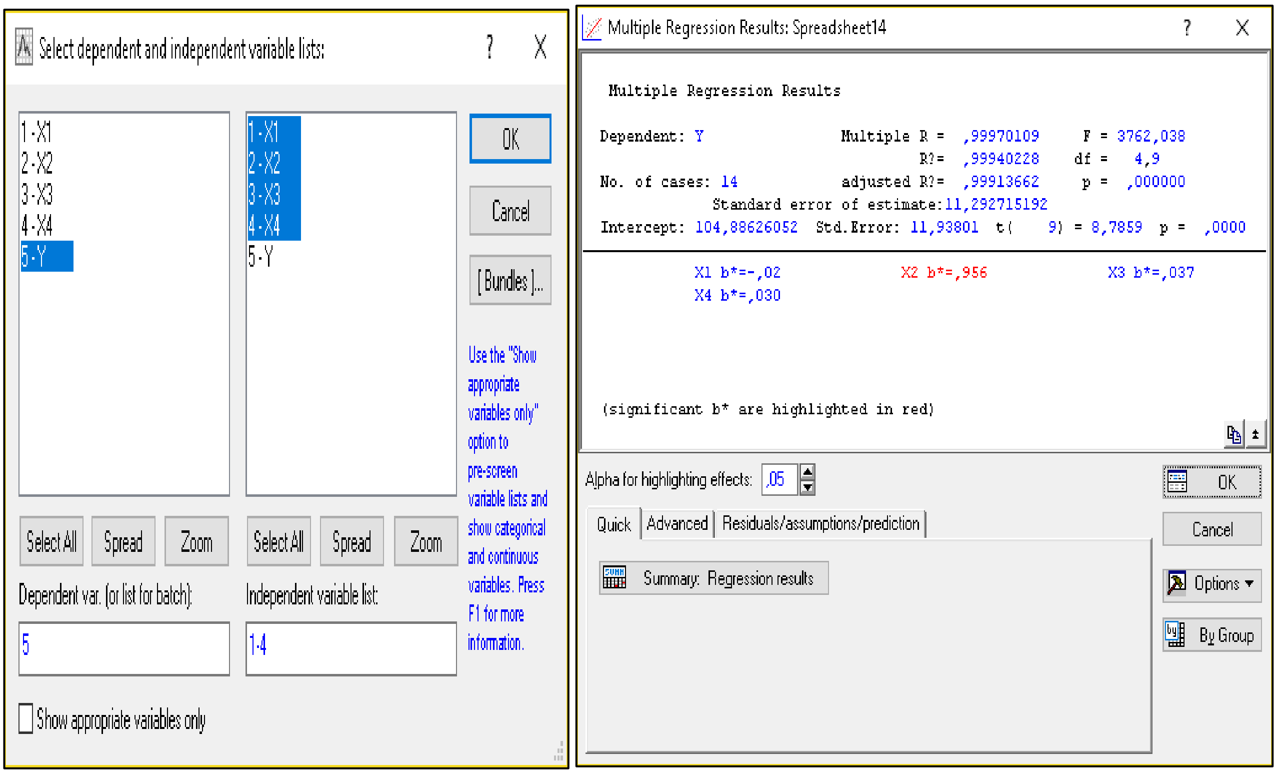


Fig. 2.16. **The stages of building a multifactor model (the end)**

2. In order to determine the parameters and quality of the model, you must initiate the *Summary: Regression results* button. The results of the calculations are shown in Fig. 2.17.

N=14	b*	Std.Err. of b*	b	Std.Err. of b	t(9)	p-value
	Regression Summary for Dependent Variable: Y (Spreadsheet14) R= ,99970109 R?= ,99940228 Adjusted R?= ,99913662 F(4,9)=3762,0 p<,00000 Std.Error of estimate: 11,293					
Intercept			104,8863	11,93802	8,78590	0,000010
X1	-0,022602	0,073626	-0,0000	0,00004	-0,30699	0,765837
X2	0,955608	0,082814	0,0033	0,00029	11,53916	0,000001
X3	0,036765	0,111982	0,0132	0,04019	0,32831	0,750184
X4	0,030262	0,081871	0,0008	0,00205	0,36963	0,720203

Fig. 2.17. **Regression results**

The results can be interpreted as follows:

1) the coefficient of the multiple correlation is equal to 0.997 (R). Since the value of the coefficient is strongly approximated to 1, we can speak about the adequacy of the model;

2) the model's determination coefficient is 0.999 (R^2). This ratio shows how much the data obtained using the model corresponds to the real data. Since the coefficient is close to 1, the adequacy of the model is confirmed;

3) the adjusted determination coefficient of the number of observations and the number of parameters is equal to 0.999 (adjusted R^2);

4) the Fisher criterion adequacy F (4, 9) = 3762 (the obtained value is more than the tabular one, which confirms the adequacy of the model);

5) the mean square error of the model is 11.93;

6) the vector of model parameters has the following form B (a_0, a_1, a_2, a_3, a_4) = (104.88; -0.001; 0.033; 0.00132; 0.0008). Thus you can form a general view of the model:

$$Y = 104.88 - 0.001X_1 + 0.033X_2 + 0.0000132X_3 + 0.0008X_4;$$

7) the vector values of the Student's criterion t (9) = (8.79; -0.3; 11.54; 0.37), proves the significance of the model parameters.

Based on the analysis of the obtained results, one can say that this model is generally adequate and qualitative, but the model parameters for variables X1, X3, and X4 are not significant, because the significance level of p is greater than 0.005.

To determine the mean and average deviations of the samples of all variables in the error analysis menu, we initiate *Descriptive statistics / Means & Standard deviations* (Fig. 2.18).

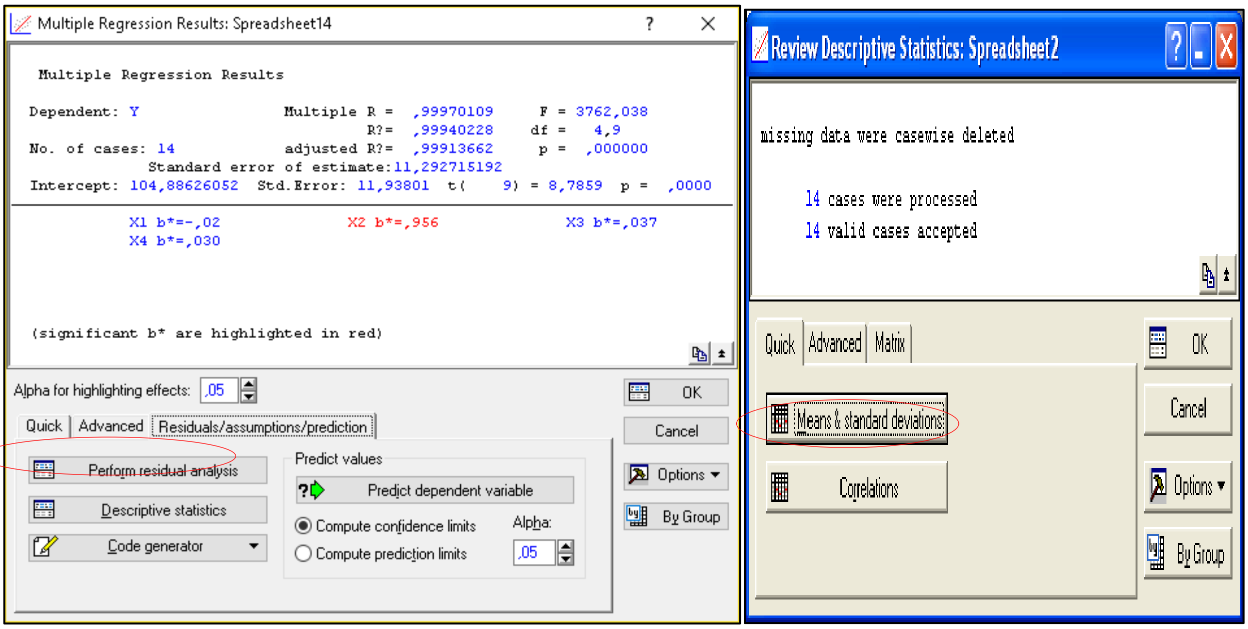


Fig. 2.18. **The *Descriptive statistics* mode**

As a result of the calculations, the average values, the mean square deviation for exogenous and endogenous variables were estimated (Fig. 2.19).

	Means and Standard Deviations (Spreadshe		
Variable	Means	Std.Dev.	N
x1	114219	805422,	14
x2	12665	111015,	14
x3	1467	1070,	14
x4	1617	15342,	14
y	543	384,	14

Fig. 2.19. **The results of the analysis**

Under the multifactor model, the error analysis is performed according to the same algorithm as for the one-factor model. Fig. 2.20 shows the results of the calculation of theoretical values for the model and model errors.

	Predicted & Residual Values (Spreadsheet14) Dependent variable: Y								
Case No.	**Observed Value**	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	186,500	173,134	13,3661	-0,963169	1,18360	6,29854	3,11557	19,4017	0,183652
2	192,500	182,076	10,4238	-0,939895	0,92306	5,16409	1,78996	13,1800	0,056971
3	198,900	205,618	-6,7182	-0,878621	-0,59492	4,72080	1,34327	-8,1409	0,018164
4	221,600	224,725	-3,1250	-0,828891	-0,27673	4,13811	0,81705	-3,6097	0,002744
5	225,800	242,896	-17,0961	-0,781596	-1,51391	4,01566	0,71527	-19,5708	0,075957
6	267,300	277,636	-10,3359	-0,691177	-0,91527	3,76245	0,51450	-11,6266	0,023533
7	345,100	336,712	8,3879	-0,537417	0,74277	5,04079	1,66169	10,4750	0,034288
8	441,500	429,298	12,2016	-0,296438	1,08048	8,01382	5,61819	24,5800	0,477175
9	544,200	553,886	-9,6861	0,027832	-0,85773	4,22389	0,89017	-11,2616	0,027827
10	720,700	716,740	3,9596	0,451700	0,35063	9,31109	7,90930	12,3674	0,163078
11	948,100	955,920	-7,8199	1,074223	-0,69248	7,22370	4,39087	-13,2359	0,112425
12	913,300	911,181	2,1189	0,957780	0,18763	9,15263	7,61104	6,1756	0,039291
13	1082,600	1079,278	3,3220	1,395293	0,29417	7,66353	5,05836	6,1580	0,027389
14	1316,600	1315,599	1,0013	2,010375	0,08867	10,61817	10,56476	8,6399	0,103503
Minimum	186,500	173,134	-17,0961	-0,963169	-1,51391	3,76245	0,51450	-19,5708	0,002744
Maximum	1316,600	1315,599	13,3661	2,010375	1,18360	10,61817	10,56476	24,5800	0,477175
Mean	543,193	543,193	-0,0000	-0,000000	-0,00000	6,38195	3,71429	2,3951	0,096143
Median	393,300	383,005	1,5601	-0,416928	0,13815	5,73132	2,45276	6,1668	0,048131

Fig. 2.20. **The results of the calculation of the model errors**

The greatest value of the error model is observed in 2009. One can conclude that during this period, the development of the country's economy significantly differed from the entire analyzed period.

Fig. 2.21 shows the modeling error distribution area.

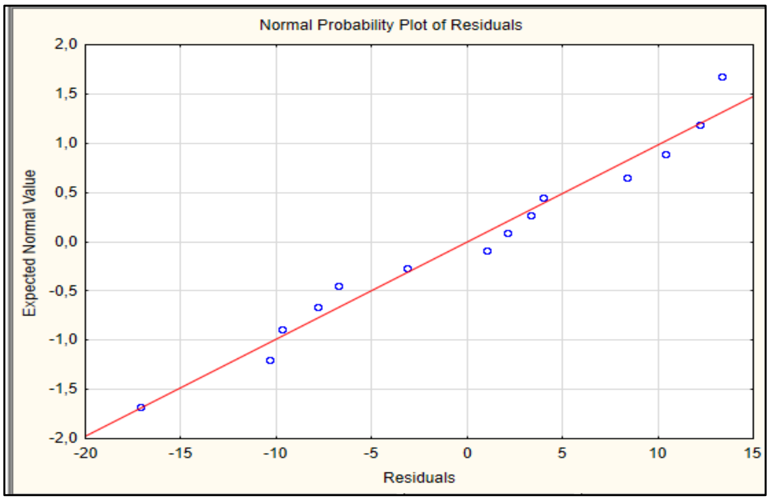


Fig. 2.21. **The modeling error distribution**

Fig. 2.22 shows the pattern distribution of the model.

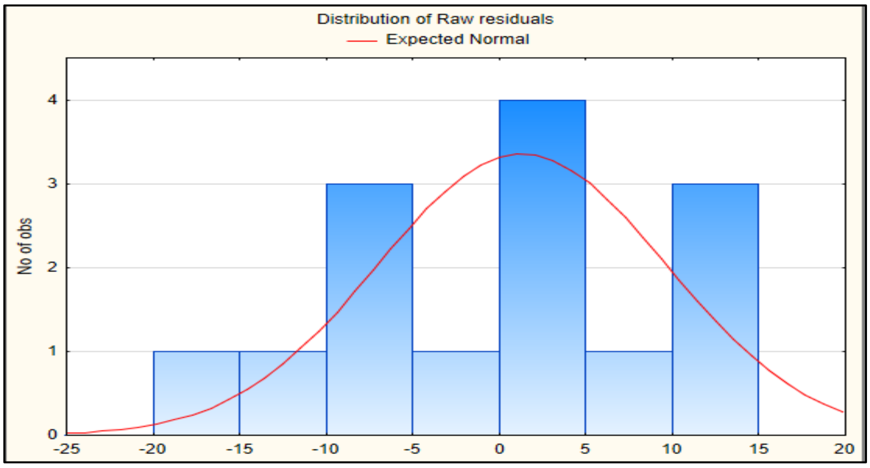


Fig. 2.22. **The histogram of the model with the distribution curve**

The calculations given in the graphs prove the hypothesis of the normal distribution law of model errors.

Conclusion. The performed calculations show that the model is adequate if the value of X1 changes by 1, the value of Y decreases by an average of 0.0001; when changing the value of X2 by 1, the value of Y increases on average by 0.0033; when the X3 value changes by 1, the Y value increases by an average of 0.0132; when the X4 value changes by 1, the Y value increases by an average of 0.008. Due to the influence of factors unaccounted for in the model, Y decreases by an average of 104.89, but most of the model parameters are insignificant. Thus, we can assume that there is multicollinearity in the model, thus the forecast is feasible only after it is eliminated.

Laboratory work 3
Model verification for multicollinearity and the algorithm for elimination of multicollinearity

The purpose of the work is to master the theoretical and practical aspects of the topic, to gain the skills in testing the econometric models for multicollinearity.

The task is to check multicollinearity in the model and eliminate it.

Guidelines

Using the data from laboratory work 2 (Table 2.2), let's check the econometric model for multicollinearity and eliminate it.

1. For comprehensive verification of the model for multicollinearity, it is expedient to use the Ferrara – Globe algorithm. All calculations by the algorithm should be done in the package MS Excel.

1.1. The first step of the algorithm is the normalization of the output data by formula (3.1):

$$z_i = \frac{x_i - \overline{x}_i}{\sigma}, \qquad (3.1)$$

where x_i is the value of the indicator i;

$\overline{x}_i$ is the arithmetic mean value of the indicator i;

σ is a standard deviation of the indicator i.

As a result, we obtain a matrix of the normalized data (Tables 3.1, 3.2).

Table 3.1

Calculation of average and standard deviation

Years	Production output, thou UAH (X1)	Volume of retail turnover of enterprises (legal entities), mln (X2)	Average cost per month per household, UAH (X3)	Direct investments, million dollars (X4)	GDP, billion UAH (У)
1	2	3	4	5	6
2005	226358	19317	395.6	2063.6	186.5
2006	356842	22151	426.5	2810.7	192.5
2007	373893	28757	541.3	3281.8	198.9
2008	460520	34417	607	3875	221.6

Table 3.1 (the end)

1	2	3	4	5	6
2009	504008	39691	658.3	4555.3	225.8
2010	603704	49994	736.8	5471.8	267.3
2011	809988	67556	903.5	6794.4	345.1
2012	995630	94332	1229.4	9047	441.5
2013	1182179	129952	1442.8	16890	544.2
2014	1565055	178233	1722	21607.3	720.7
2015	2072172	246903	2590.4	29542.7	948.1
2016	1955685	230955	2754.1	35616.4	913.3
2017	2388289	280890	3072.7	40053	1082.6
2018	2496365	350059	3456	44806	1316.6
X average	1142192	126657.6429	1466.885714	16172.5	543.1928571
Standard deviation	805422.14	111015.7743	1070.833261	15342.796	384.3245513

Table 3.2

The normalized values of the model output

Years	X1	X2	X3	X4	Y
2005	-1.137086	-0.966895	-1.000423	-0.919578	-0.928103
2006	-0.975079	-0.941368	-0.971566	-0.870884	-0.912491
2007	-0.953908	-0.881862	-0.864360	-0.840179	-0.895839
2008	-0.846354	-0.830879	-0.803006	-0.801516	-0.836774
2009	-0.792360	-0.783372	-0.755100	-0.757176	-0.825846
2010	-0.668579	-0.690565	-0.681792	-0.697441	-0.717864
2011	-0.412459	-0.532372	-0.526119	-0.611238	-0.515431
2012	-0.181969	-0.291181	-0.221777	-0.464420	-0.264602
2013	0.049647	0.029675	-0.022492	0.046765	0.002621
2014	0.525020	0.464577	0.238239	0.354225	0.461868
2015	1.154649	1.083138	1.049196	0.871432	1.053555
2016	1.010021	0.939482	1.202068	1.267298	0.963007
2017	1.547135	1.389283	1.499593	1.556463	1.403520
2018	1.681321	2.012339	1.857539	1.866250	2.012380

With the help of the built-in CORREL function it is necessary to calculate the matrix of pair correlations on the normalized data.

The matrix of correlation coefficients in Excel is constructed using the *Correlation* tool from the *Data analysis* package.

On the *Data* tab in the *Analysis* group, open the *Data analysis* package. If the button is not available, you need to add it (*Excel Options – Add-ins*). In the analysis tools list, select *Correlation* (Fig. 3.1).

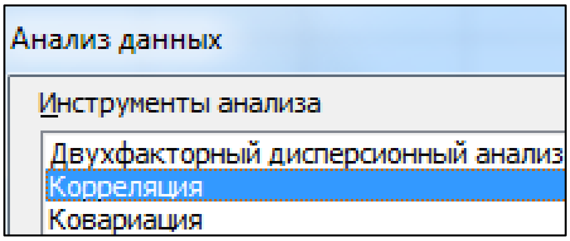


Fig. 3.1. **Selection of the option *Correlation***

Click OK. Set the parameters for data analysis. The input interval is the range of cells with values. Grouping by columns is used (analyzed data are grouped into columns) (Fig. 3.2). The output interval is a reference to the cell from which the matrix is to be built. The size of the range will be determined automatically.

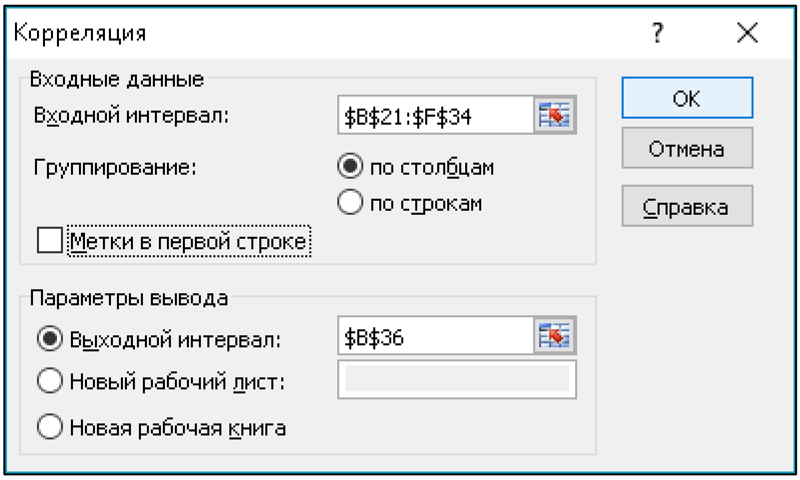


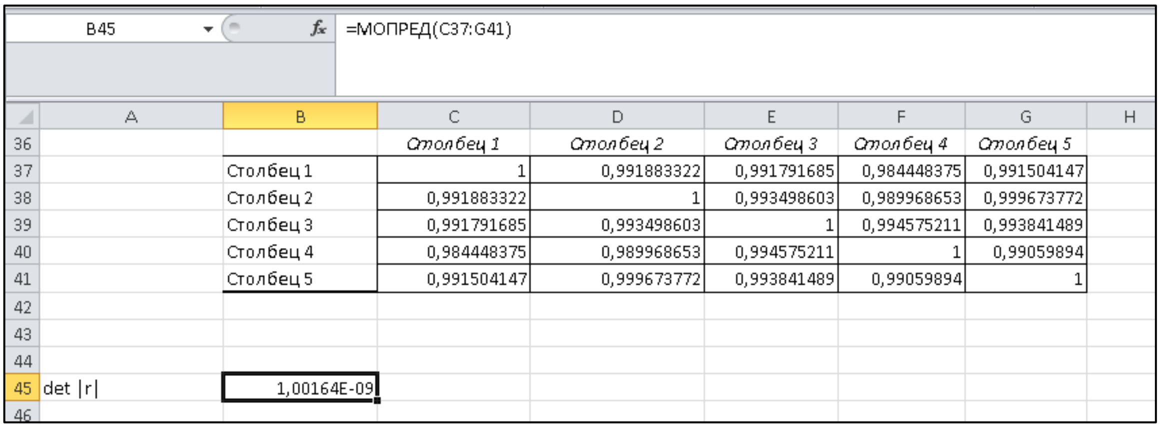
Fig. 3.2. **Choosing the parameters for data analysis**

The calculated matrix has the following form:

	Column 1	*Column 2*	*Column 3*	*Column 4*	*Column 5*
Column 1	1	0.991883322	0.991791685	0.984448375	0.991504147
Column 2	0.991883322	1	0.993498603	0.989968653	0.999673772
Column 3	0.991791685	0.993498603	1	0.994575211	0.993841489
Column 4	0.984448375	0.989968653	0.994575211	1	0.99059894
Column 5	0.991504147	0.999673772	0.993841489	0.99059894	1

A strong direct relationship is found between the values of Y and X_1, X_2, X_3, X_4. The same connection is also found between all factors.

The next step is to find the determinant of the correlation matrix r. The function *МОПРЕД* returns the matrix determinant. The matrix is written in an array (Fig. 3.3).



B45 =МОПРЕД(C37:G41)

	A	B	C	D	E	F	G	H
36			Столбец 1	Столбец 2	Столбец 3	Столбец 4	Столбец 5	
37		Столбец 1	1	0,991883322	0,991791685	0,984448375	0,991504147	
38		Столбец 2	0,991883322	1	0,993498603	0,989968653	0,999673772	
39		Столбец 3	0,991791685	0,993498603	1	0,994575211	0,993841489	
40		Столбец 4	0,984448375	0,989968653	0,994575211	1	0,99059894	
41		Столбец 5	0,991504147	0,999673772	0,993841489	0,99059894	1	
42								
43								
44								
45	det |r|	1,00164E-09						
46								

Fig. 3.3. **The determinant of the correlation matrix r**

1.2. Analysis of general multicollinearity of the model is conducted using the Pearson criterion or criterion X^2:

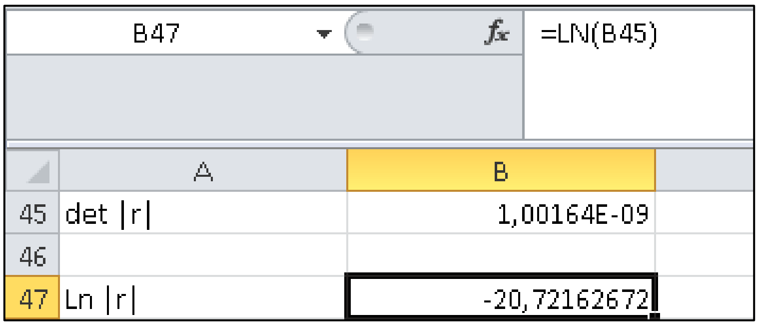
$$X^2 = -\left[n - 1 - \frac{1}{6}(2m + 5)\right] Ln|r|, \tag{3.2}$$

where n is the number of observations;

m is the number of independent variables;

Ln is the natural logarithm of the number.

To find the natural algorithm ($Ln|r|$) you must use the built-in function MS Excel *Ln*, which returns the natural algorithm of the number (Fig. 3.4).



B47 =LN(B45)

	A	B
45	det |r|	1,00164E-09
46		
47	Ln |r|	-20,72162672

Fig. 3.4. **Finding the natural algorithm**

$$X^2 = -\left[14 - 1 - \frac{1}{6}(2 \times 4 + 5)\right](-20.722) = 224.4$$

At a degree of freedom 1 / 2m (m - 1) = 1/2 × 4 (4 - 1) = 6 and the levels of significance a = 0.05 the tabular (critical value) X^2_{table} is 12.592:

$X^2_{table}(\alpha = 0.05;\ k = 6) = 12.592 => |\ X^2_{calc}| > X^2_{table}$.

So, the model is characterized by general multicollinearity.

1.3. It is necessary to evaluate the Fisher coefficient by formula (3.3):

$$F_k = (c_{kk} - 1) \times \frac{n - m}{m - 1}, \qquad (3.3)$$

where F is the Fisher coefficient;

c_{kk} is diagonal elements of the matrix C.

The matrix C (the dimension of the matrix is 5X5), inversed to the correlation matrix r, is determined using the built-in function МОБР.

First, we need to select an array whose dimension corresponds to the dimension of the matrix r (in our case, B50:F54.) Then enter the formula and finish the input with a combination of the buttons Ctrl + Shift + Enter.

The calculation of the elements of the matrix C is shown in Fig. 3.5.

B50 {=МОБР(C37:G41)}

	A	B	C	D	E	F	G
36			*Столбец 1*	*Столбец 2*	*Столбец 3*	*Столбец 4*	*Столбец 5*
37		Столбец 1	1	0,991883322	0,991791685	0,984448375	0,991504147
38		Столбец 2	0,991883322	1	0,993498603	0,989968653	0,999673772
39		Столбец 3	0,991791685	0,993498603	1	0,994575211	0,993841489
40		Столбец 4	0,984448375	0,989968653	0,994575211	1	0,99059894
41		Столбец 5	0,991504147	0,999673772	0,993841489	0,99059894	1
42							
43							
44							
45	det |r|	1,00164E-09					
46							
47	Ln |r|	-20,72162672					
48							
49							
50		82,47537938	-80,53925274	-60,54329354	21,29509785	37,81383357	
51	C=R^(-1)=	-80,53925274	1631,040598	25,88752305	22,57836838	-1598,747707	
52		-60,54329354	25,88752305	191,0777009	-95,13659361	-61,50888924	
53		21,29509785	22,57836838	-95,13659361	102,4580169	-50,62928959	
54		37,81383357	-1598,747707	-61,50888924	-50,62928959	1673,016985	
55							

Fig. 3.5. **The matrix C**

Using the diagonal elements of the matrix C, we calculate the F-criterion for each independent variable by the formulas:

$$F_1 = (c_{11} - 1) \times \frac{n-m}{m-1} = (82.475 - 1) \times 14 - 4/4 - 1 = 271.58,$$

$$F_2 = (c_{22} - 1) \times \frac{n-m}{m-1} = (1631.041 - 1) \times 14 - 4/4 - 1 = 5432.93,$$

$$F_3 = (c_{33} - 1) \times \frac{n-m}{m-1} = (191.078 - 1) \times 14 - 4/4 - 1 = 633.53,$$

$$F_4 = (c_{44} - 1) \times \frac{n-m}{m-1} = (102.458 - 1) \times 14 - 4/4 - 1 = 338.16.$$

Each of the obtained Fisher coefficients substantially exceeds the table value for α = 0.05, $k_1 = n - m$ = 14 – 4 = 10, $k_2 = m - 1$ = 4 – 1 = 3, which is F = 8.78. If the calculated value exceeds the table one, then k is multi-collinear, so we see that the variables of the model cause multicollinearity.

1.4. Student's coefficient is used to determine the pairwise multi-collinearity.

Next, we will determine the partial correlation coefficients, which characterize the tightness of the relationship between the two variables, provided that the other variables are not affected.

Using the matrix C we calculate the partial correlation coefficients (formula 3.4):

$$r_{kj} = \frac{-c_{kj}}{\sqrt{c_{kk} \times c_{jj}}}, \tag{3.4}$$

where C_{kj} is an element of the matrix C located in the corresponding k-row and j-column;

C_{kk} and C_{jj} are the diagonal matrix elements.

$$r_{12} = \frac{-c_{kj}}{\sqrt{c_{kk} \times c_{jj}}} = -(-80.539)/\sqrt{82.475 \times 1631.041} = 0.2196,$$

$$r_{13} = 0.482,$$

$$r_{14} = -0.232,$$

$$r_{23} = -0.046,$$

$$r_{24} = -0.055,$$

$$r_{34} = 0.68.$$

The partial correlation coefficients characterize the closeness of the relationship between the two variables, provided that the third one does not affect this relationship.

1.5. On the basis of the found partial correlation coefficients, we find the estimated values of Student's t-criterion:

$$t_{kj} = r_{kj} \times \frac{\sqrt{n-m}}{\sqrt{1-r^2}}, \qquad (3.5)$$

$t_{12} =$ 0.7118,
$t_{13} =$ 1.7409,
$t_{14} =$ -0.7530,
$t_{23} =$ -0.1468,
$t_{24} =$ -0.1749,
$t_{34} =$ 2.9323.

In order to conclude on the presence of multicollinearity, it is necessary to compare the obtained values with the tabular ones.

So, for the level of significance $a = 0.05$ at the $n - m$ degrees of freedom according to the statistical tables of the Student's t-distribution, we find the critical value of the Student's t-criterion: $t_{\text{table }(0.05;10)} = 2.23$.

If the actual value of the Student t-criterion is more than the tabular one, this indicates the existence of multiconflict between the pairs of factors.

So, it can be argued that there is multicellularity between the variables X_3 and X_4.

Thus, by analyzing the model for multicollinearity by different methods, we can conclude that the model has multicollinearity. This is due to the presence of a connection of varying degrees between different features.

2. To get rid of multicollinearity we use the methods of step-by-step inclusion and stepwise exclusion of variables.

In the package Statistica in the module *Multiple Regression*, *the Forward stepwise* method and the *Backward stepwise* method are selected in the Start menu in the menu *Advanced* (Fig. 3.6).

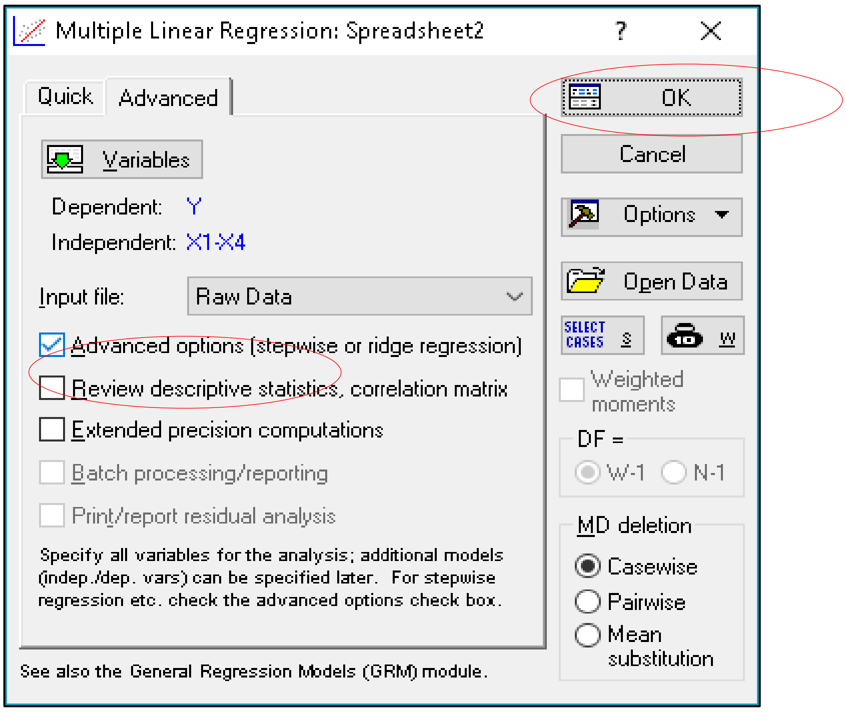


Fig. 3.6. **Choosing the parameters in the module *Multiple regression***

The first step is to select the method of stepwise exclusion of the parameters (*Backward stepwise*) (Fig. 3.7).

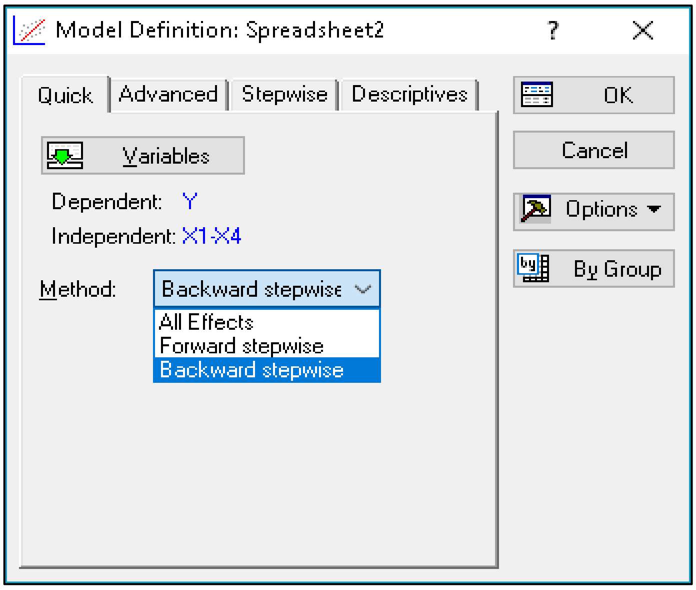


Fig. 3.6. **Choosing the method of stepwise exclusion *Backward stepwise***

Fig. 3.7 shows the results of building a multifactor model by a step-by-step exclusion method.

	Regression Summary for Dependent Variable: Y (Spreadsheet2) R= ,99967377 R?= ,99934765 Adjusted R?= ,99929329 F(1,12)=18383, p<0,0000 Std.Error of estimate: 10,217					
N=14	**b***	Std.Err. of b*	b	Std.Err. of b	t(12)	p-value
Intercept			104,8609	4,231761	24,7795	0,000000
X2	0,999674	0,007373	0,0035	0,000026	135,5842	0,000000

Fig. 3.7**. The results of the analysis**

The implementation of the step-by-step method is carried out by selecting the appropriate menu item. As a result of the implementation of the calculation algorithms, we obtain the final form of the econometric model:

$$Y = 104.861 + 0.0035\ X_2.$$

The analysis of all built models make it possible to conclude that there is multicollinearity (linear dependence) caused by errors in the specification, therefore, it is advisable to carry out additional analysis of the model, or to use the methods of curtailing the sign space.

An important prerequisite for building a qualitative regression model using the least squares method is the independence of the values of random deviations ε_i from the values of deviations in all other observations. The absence of dependence guarantees the absence of correlation between any deviations ($\sigma(\varepsilon_i, \varepsilon_j) = \mathrm{cov}(\varepsilon_i, \varepsilon_j = 0 \text{ if } i \neq j$) and, in particular, between adjacent deviations ($\sigma(\varepsilon_{i-1}, \varepsilon_i) = 0$), i =2, 3, ..., n.

Autocorrelation (sequential correlation) is defined as the correlation between the observed indicators arranged in time (time series) or in space (cross data). Autocorrelation of residues (deviations) is commonly found in regression analysis when using time series data. If cross data are used, the presence of autocorrelation (spatial correlation) is extremely rare.

Autocorrelation does not prevent finding of the relationship between the investigated index and the factors affecting it, as well as the determination of the parameters and statistical characteristics of the equation, but its presence does not guarantee the reliability of the regression equation and the parameters and the possibility of building confidence intervals. Consequently, autocorrelation is more "dangerous" when making a forecast than when conducting economic analysis.

The consequences of autocorrelation are somewhat similar to those of heteroscedasticity. Among them, the following are usually distinguished.

1. Estimates of parameters, while remaining linear and unshakable, cease to be effective. Consequently, they cease to have the properties of the best linear immutable estimates (BLIE ratings).

2. Dispersion estimates are biased. Most often, dispersion, calculated by standard formulas, is understated, which leads to an increase in t-statistic. This can lead to the recognition of statistically significant explanatory variables, which in reality may not be.

3. The estimation of the regression dispersion $S^2 = \sum \frac{e_t^2}{n-m-1}$ is a biased estimate of the actual value σ^2, which in many cases is underscored.

4. In view of the foregoing conclusions, the t- and F-statistics that determine the significance of the regression coefficients and the determination coefficient may not be valid. As a result, the predicted quality of the model deteriorates.

Among the main reasons that cause the appearance of autocorrelation, one can distinguish:

1) specifications;
2) inertia;
3) the effect of the web;
4) smoothing the data.

The presence of autocorrelation is determined by the following criteria:

- the Darwin – Watson criterion;
- the von Neumann criteria;
- the non-cyclic criterion of autocorrelation;
- the cyclic criterion of autocorrelation.

For example, to calculate the Darwin – Watson criterion (statistics), it is possible to use the package Statistica 8.0, the module *Multiple Regression*. Then choose the tab *Residals/assumptions/prediction*, initiate the *Perform residual analysis* button, choose the tab *Advanced* and click the button *Durbin – Watson statistic*.

The calculated value of the DW statistics is compared with its critical value, which cannot be completely determined due to its dependence on regressor values, and with the lower (d_l) and upper (d_u) limits of the critical value (d_{table}): $d_l \le d_{table} \le d_u$.

The boundaries d_u and d_l are selected according to the number of observations (n), the number of regressors (k) and the level of significance (α).

The values of d_u and d_l are determined by the table of critical values of Durbin – Watson and conclusions are drawn based on the following scheme:

1) if $0 < DW < d_l$, there is a positive autocorrelation of the residues;

2) if $d_l \leq DW \leq d_u$, the conclusion about the presence of autocorrelation is not determined (the zone of uncertainty);

3) if $d_u < DW < 4 - d_u$, there is no autocorrelation;

4) if $4 - d_u, \leq DW \leq 4 - d_l$, the conclusion about the presence of autocorrelation is not determined;

5) if $4 - d_l < DW < 4$, there is a negative autocorrelation of the residues.

Topic 3. Modeling and forecasting the development of trends

Laboratory work 4
Building a time series decomposition model

The purpose of the work is to master the theoretical and practical aspects of the topic, to gain the skills in building decomposition models.

The task is to determine the form of the decomposition model, to identify all the components, to forecast the trend component, to carry out spectral analysis of the cyclic constituent and the composition of the model and check its quality.

Guidelines

It is necessary to form a dynamic series and present it as a file in the package Statistica 8.0 (Fig. 4.1).

Data: Spreadsheet2* (2v by 32c)

	1 T	2 GDP
1	1	189028
2	2	214103
3	3	250306
4	4	259908
5	5	217074
6	6	255545
7	7	300446
8	8	306281
9	9	258591
10	10	310277
11	11	368488
12	12	362635
13	13	292324
14	14	346005
15	15	387109
16	16	379231
17	17	303753
18	18	354814
19	19	398000
20	20	408631
21	21	316905
22	22	382391
23	23	440476
24	24	447143
25	25	375991
26	26	456715
27	27	566997
28	28	588841
29	29	455637
30	30	535324
31	31	669170
32	32	723051

Fig. 4.1. **The initial data in Statistica 8.0**

Suppose, we know the monthly dynamics of Ukraine's GDP (in million UAH), which is presented in Table 4.1.

Table 4.1

The input data

Period (T)	The GDP volume of Ukraine
1st quarter 2011	189 028
2nd quarter	214 103
3rd quarter	250 306
4th quarter	259 908
1st quarter 2012	217 074
2nd quarter	255 545
3rd quarter	300 446
4th quarter	306 281
1st quarter 2013	258 591
2nd quarter	310 277
3rd quarter	368 488
4th quarter	362 635
1st quarter 2014	292 324
2nd quarter	346 005
3rd quarter	387 109
4th quarter	379 231
1st quarter 2015	303 753
2nd quarter	354 814
3rd quarter	398 000
4th quarter	408 631
1st quarter 2016	316 905
2nd quarter	382 391
3rd quarter	440 476
4th quarter	447 143
1st quarter 2017	375 991
2nd quarter	456 715
3rd quarter	566 997
4th quarter	588 841
1st quarter 2018	455 637
2nd quarter	535 324
3rd quarter	669 170
4th quarter	723 051

In order to determine the model of decomposition of time series components (additive or multiplicative), we present the initial data in the form of a graph (Fig. 4.2).

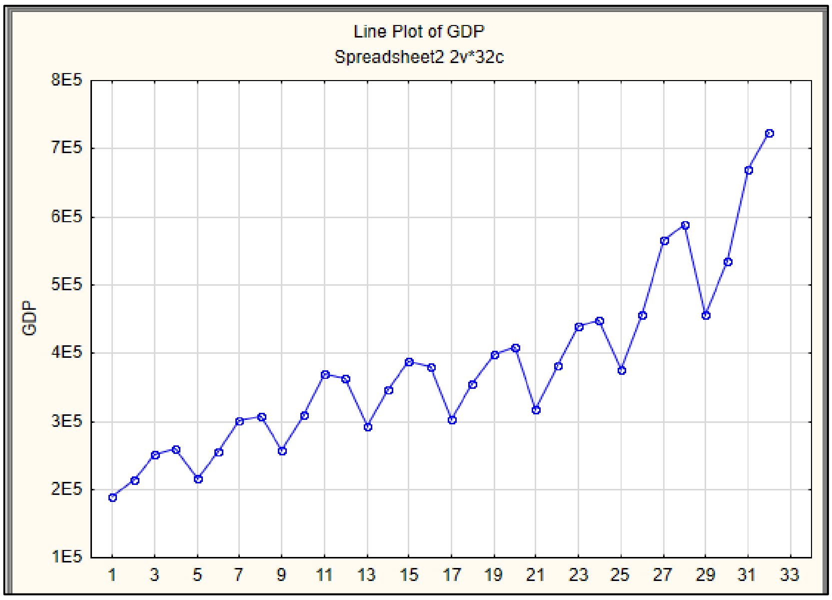


Fig. 4.2. **The line plot of the GDP**

Visual analysis shows the presence of a positive trend and some seasonality. Since the volume of the GDP does not have a constant clearly pronounced tendency to increase or decrease the amplitude of the values, it is recommended that the multiplicative time series model be used, which is generally given by formula (4.1):

$$Y = T_t \times S_t \times C_t \times I, \qquad (4.1)$$

where T is a trend component;

C is a cyclic component;

S is a seasonal component;

I is a random component.

In the case of a constant amplitude of changes in the values of the time series, it is expedient to use the additive model.

To determine the presence of seasonality and the values of the seasonal lag, we use the *Time series analysis* module (Fig. 4.3).

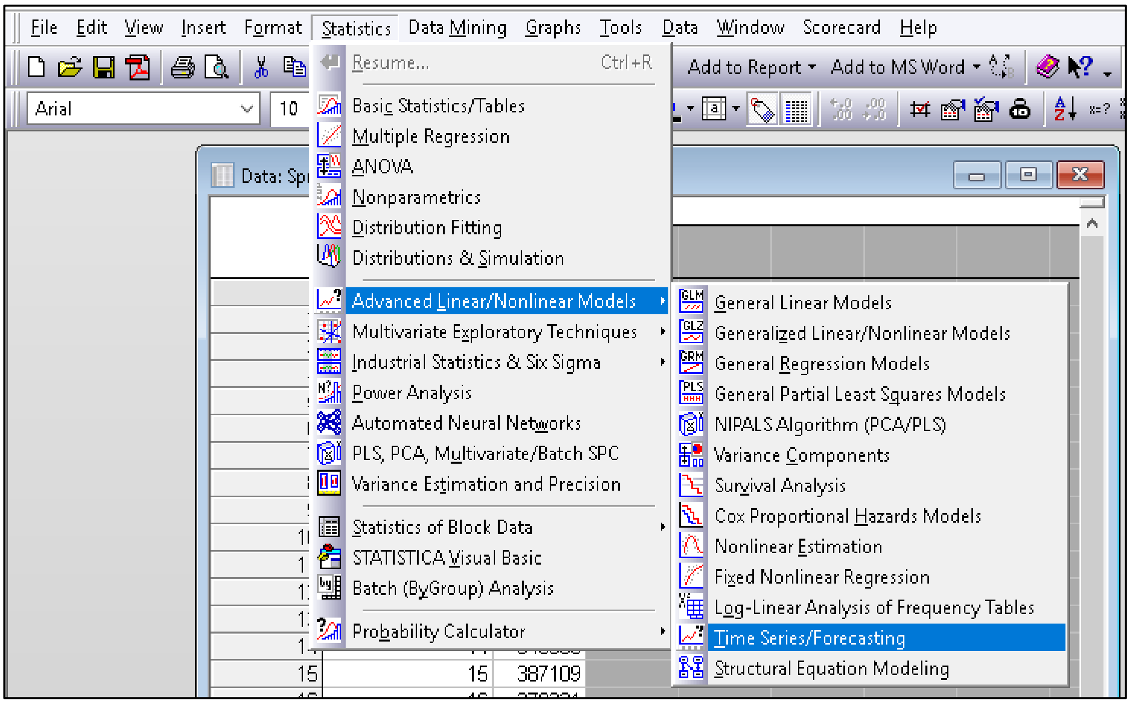


Fig. 4.3. **The *Time series analysis* module in Statistica 8.0**

Then, we need to choose a variable for analysis – the GDP (Fig. 4.4).

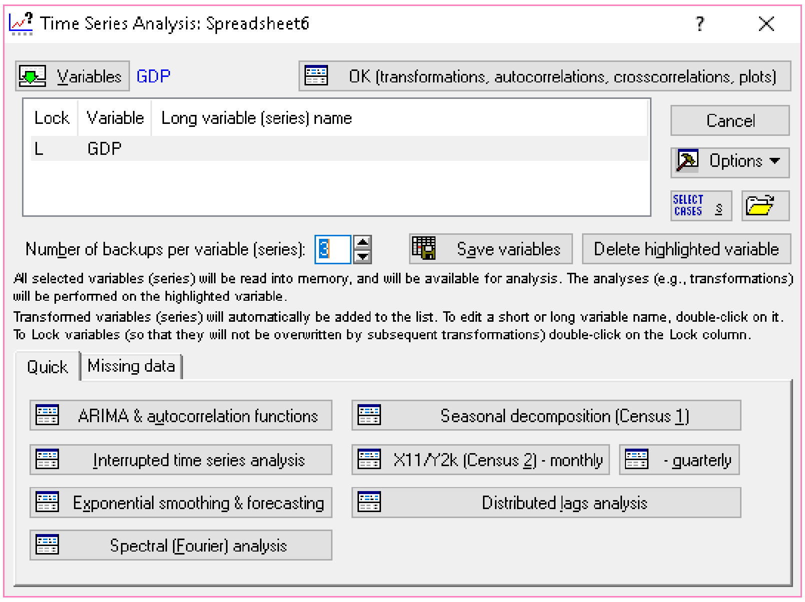


Fig. 4.4. **Choosing the parameters in the *Time series analysis* module**

The dialog box for this analysis contains the area where the original and converted time series are stored. The number of copies per row can be set by the user alone (a minimum of 3 is recommended). Your source row is denoted by the locked variable L (LOCK). This means that it will always be saved and will not be deleted after all the manipulations.

Confirmations of visual analysis will be performed analytically, namely through:

1) autocorrelation analysis;

2) Fourier analysis.

Autocorrelation analysis helps to identify seasonality and determine the seasonal lags of the time series. To do this, select the *Autocorrelation* tab (Fig. 4.5).

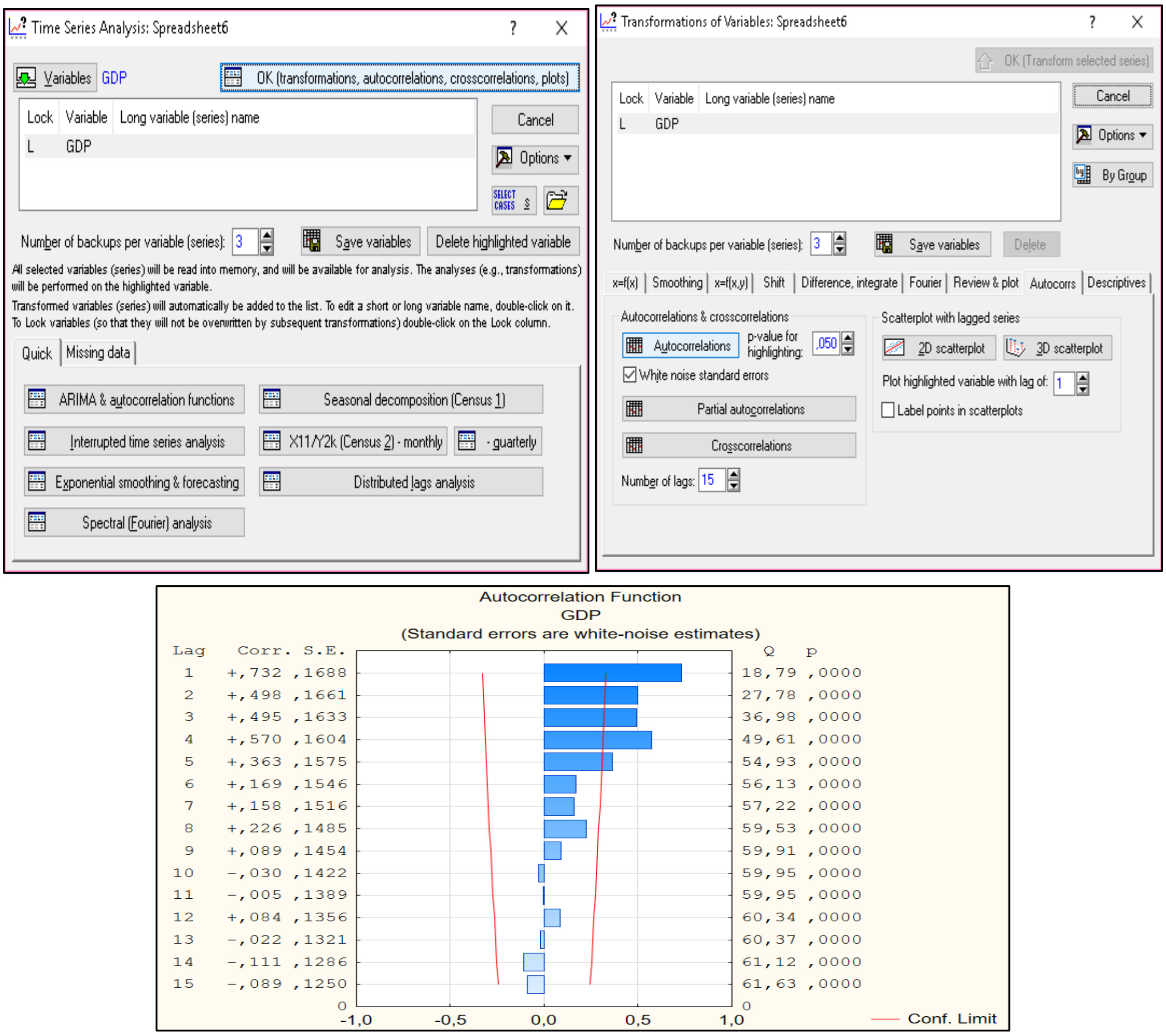


Fig. 4.5. **The steps of the autocorrelation analysis**

As we see in Fig. 4.5, the greatest correlation values fall on the 1st lag, then there is a decline and its maximum autocorrelation coefficient appears in 4 lags. So, in our data there is a trend and seasonality equal to 4.

To confirm the presence of seasonality and show the presence of hidden seasonalities, which could not be determined using the autocorrelation function, we use the Fourier method. To do this, go back to the previous level (Fig. 4.6).

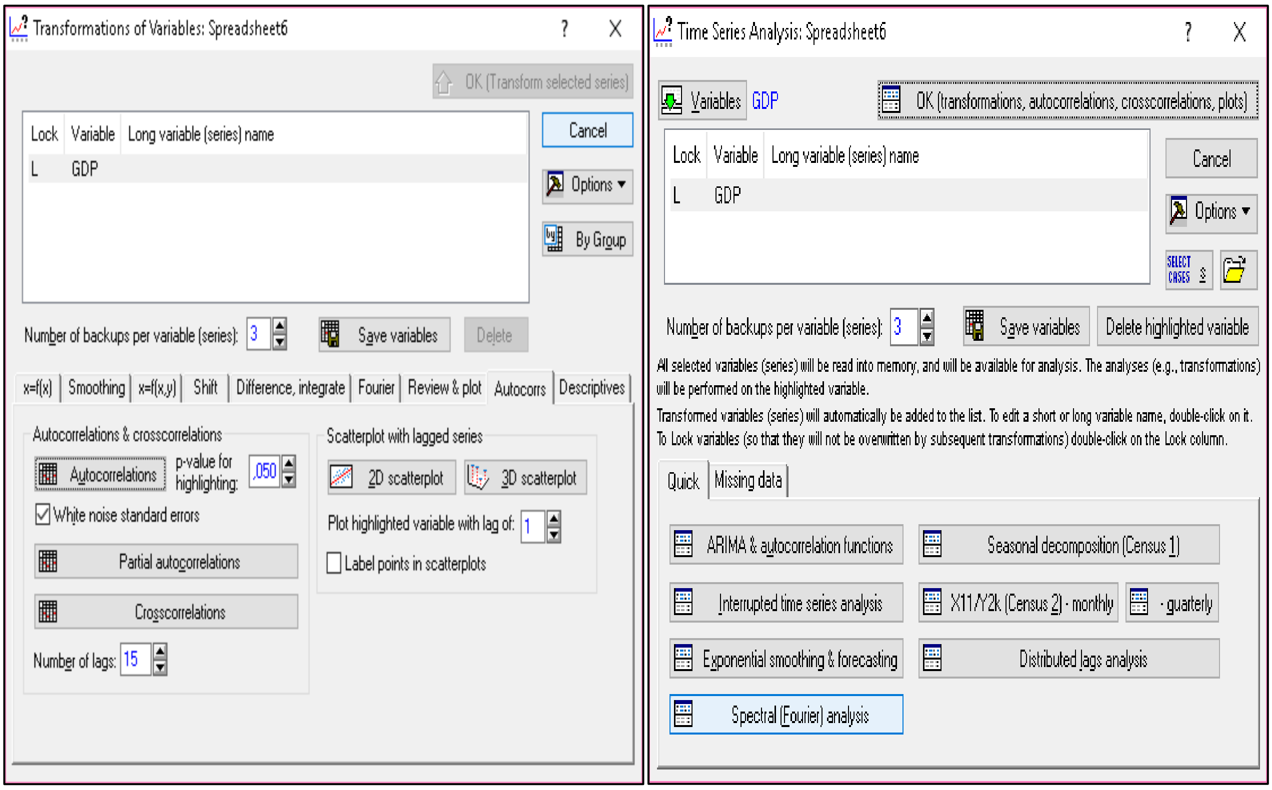


Fig. 4.6. **The Fourier method**

Here we are interested in one-dimensional (single series) analysis. On the X-axis, we plot the period and set the spectral plane (Fig. 4.7).

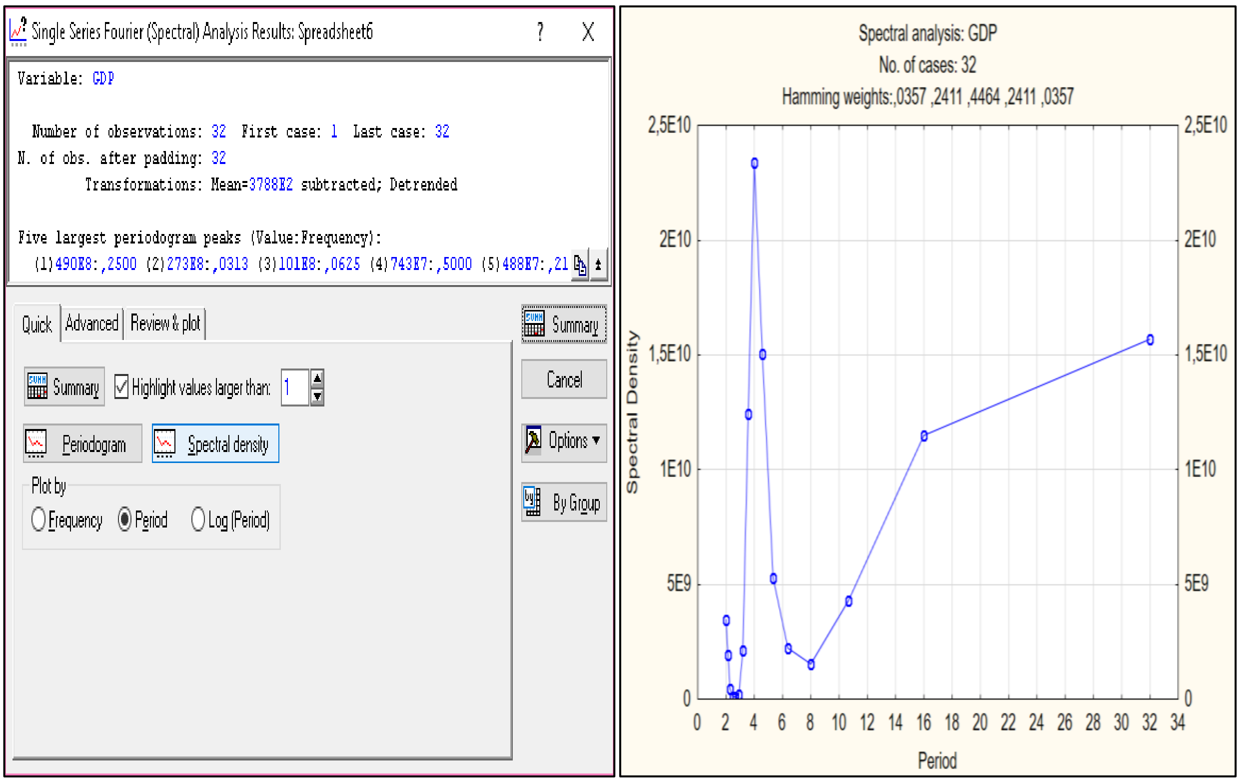


Fig. 4.7. **The results of the Fourier method**

We obtain a plot of the spectral density over the period. Obviously, the absolute maximum is reached at the point with lag 4, and there is also a seasonal component equal to half a quarter (lag 2). But since the value of the spectral density at this point is smaller, seasonality with lag 4 affects the variability of the data to a greater extent than seasonality with lag 2.

So, using the graphical and analytical methods, we get convinced of the presence of a trend-cyclic and seasonal component in the model.

Decomposition of the time series is carried out based on the following components: trend-cyclic, seasonal and random.

To do this, select the *Seasonal Decomposition* tab in the start-up panel of the *Advanced Linear / Nonlinear Models / Time Series / Forecasting* module and set the seasonal decomposition parameters (Fig. 4.8):

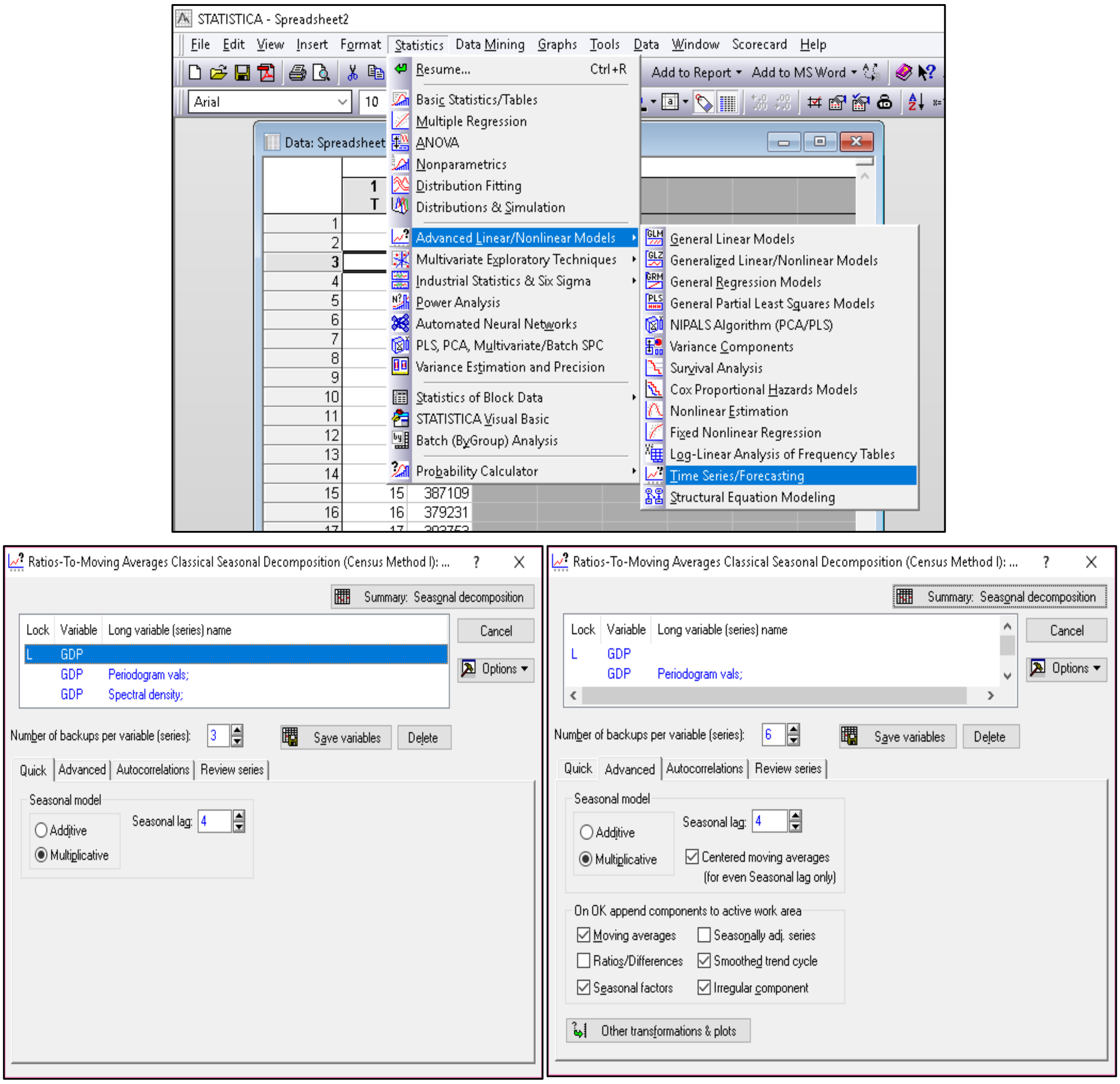


Fig. 4.8. **Choosing the *Time series analysis (TSA)***

Specifying the parameters of the seasonal decomposition model, we obtain the following result (Fig. 4.9).

	Seasonal Decomposition: Multipl. season (4); Centered means (Spreadsheet6) GDP						
Case	**GDP**	Moving Averages	Ratios	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	189028,0			84,7703	222988,6	219614,0	1,015366
2	214103,0			97,2987	220047,1	223577,9	0,984208
3	250306,0	231842,0	107,9640	109,9289	227698,0	231505,8	0,983552
4	259908,0	240528,0	108,0573	108,0021	240650,9	241353,6	0,997089
5	217074,0	251975,8	86,1488	84,7703	256073,3	252867,6	1,012677
6	255545,0	264039,9	96,7827	97,2987	262639,6	263435,9	0,996977
7	300446,0	275026,1	109,2427	109,9289	273309,3	274834,0	0,994452
8	306281,0	287057,3	106,6968	108,0021	283588,1	287668,0	0,985817
9	258591,0	302404,0	85,5118	84,7703	305049,2	303180,1	1,006165
10	310277,0	317953,5	97,5857	97,2987	318891,1	317393,1	1,004720
11	368488,0	329214,4	111,9295	109,9289	335205,7	329424,9	1,017548
12	362635,0	337897,0	107,3212	108,0021	335766,7	337988,8	0,993426
13	292324,0	344690,6	84,8076	84,7703	344842,6	344959,3	0,999662
14	346005,0	349092,8	99,1155	97,2987	355611,0	349745,3	1,016771
15	387109,0	352595,9	109,7883	109,9289	352144,8	352565,6	0,998807
16	379231,0	355125,6	106,7878	108,0021	351133,0	354957,1	0,989227
17	303753,0	357588,1	84,9449	84,7703	358324,9	357863,0	1,001291
18	354814,0	362624,5	97,8461	97,2987	364664,6	362692,9	1,005436
19	398000,0	367943,5	108,1688	109,9289	362052,1	367151,1	0,986112
20	408631,0	373034,6	109,5424	108,0021	378354,7	373835,5	1,012089
21	316905,0	381791,3	83,0048	84,7703	373839,8	380776,3	0,981783
22	382391,0	391914,8	97,5699	97,2987	393007,2	391161,4	1,004719
23	440476,0	404114,5	108,9978	109,9289	400691,6	403721,9	0,992494
24	447143,0	420790,8	106,2626	108,0021	414013,3	421434,1	0,982391
25	375991,0	445896,4	84,3225	84,7703	443541,1	445990,7	0,994508
26	456715,0	479423,8	95,2633	97,2987	469394,6	476229,1	0,985649
27	566997,0	507091,8	111,8135	109,9289	515785,1	506400,8	1,018531
28	588841,0	526873,6	111,7613	108,0021	545212,6	529086,8	1,030479
29	455637,0	549471,4	82,9228	84,7703	537496,3	547533,4	0,981668
30	535324,0	579019,3	92,4536	97,2987	550186,0	573077,9	0,960054
31	669170,0			109,9289	608729,7	609464,8	0,998794
32	723051,0			108,0021	669478,8	627658,3	1,066629

Fig. 4.9. **The results of TSA in Statistica 8.0**

At the next stage of the laboratory work, it is necessary to copy the decomposition results, namely the trend-cyclic, seasonal and random components into the window with the initial data (Fig. 4.10).

	1 T	2 GDP	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.
1	1	189028	219614,0	84,7703	1,015366
2	2	214103	223577,9	97,2987	0,984208
3	3	250306	231505,8	109,9289	0,983552
4	4	259908	241353,6	108,0021	0,997089
5	5	217074	252867,6	84,7703	1,012677
6	6	255545	263435,9	97,2987	0,996977
7	7	300446	274834,0	109,9289	0,994452
8	8	306281	287668,0	108,0021	0,985817
9	9	258591	303180,1	84,7703	1,006165
10	10	310277	317393,1	97,2987	1,004720
11	11	368488	329424,9	109,9289	1,017548
12	12	362635	337988,8	108,0021	0,993426
13	13	292324	344959,3	84,7703	0,999662
14	14	346005	349745,3	97,2987	1,016771
15	15	387109	352565,6	109,9289	0,998807
16	16	379231	354957,1	108,0021	0,989227
17	17	303753	357863,0	84,7703	1,001291
18	18	354814	362692,9	97,2987	1,005436
19	19	398000	367151,1	109,9289	0,986112
20	20	408631	373835,5	108,0021	1,012089
21	21	316905	380776,3	84,7703	0,981783
22	22	382391	391161,4	97,2987	1,004719
23	23	440476	403721,9	109,9289	0,992494
24	24	447143	421434,1	108,0021	0,982391
25	25	375991	445990,7	84,7703	0,994508
26	26	456715	476229,1	97,2987	0,985649
27	27	566997	506400,8	109,9289	1,018531
28	28	588841	529086,8	108,0021	1,030479
29	29	455637	547533,4	84,7703	0,981668
30	30	535324	573077,9	97,2987	0,960054
31	31	669170	609464,8	109,9289	0,998794
32	32	723051	627658,3	108,0021	1,066629

Fig. 4.10. **Adding the decomposition components in the file**

Next, we need to visualize the components of the composition model. For this purpose, we need to return to the *Analysis* tab, go to the charts and alternately choose the variable we need to press the *Graph* button (Fig. 4.11).

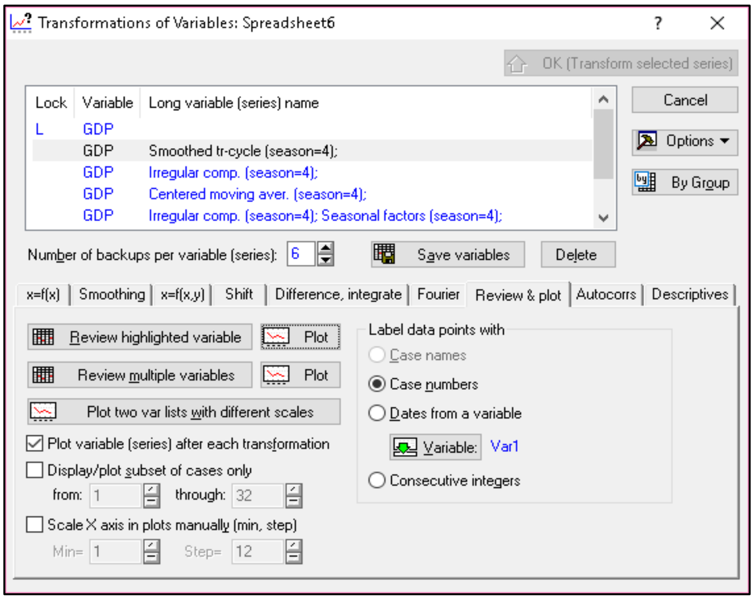


Fig. 4.11. **Choosing the parameters**

Then we need to build a trend-cycle component (Fig. 4.12).

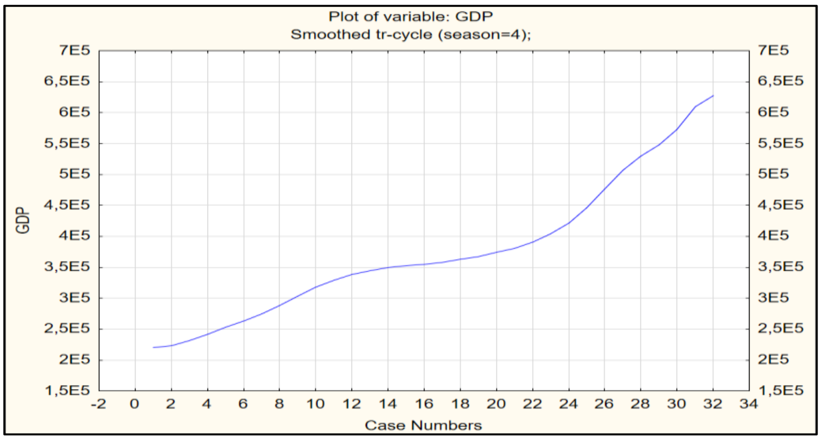


Fig. 4.12. **The trend-cycle component**

Then we need to build the seasonal component (Fig. 4.13).

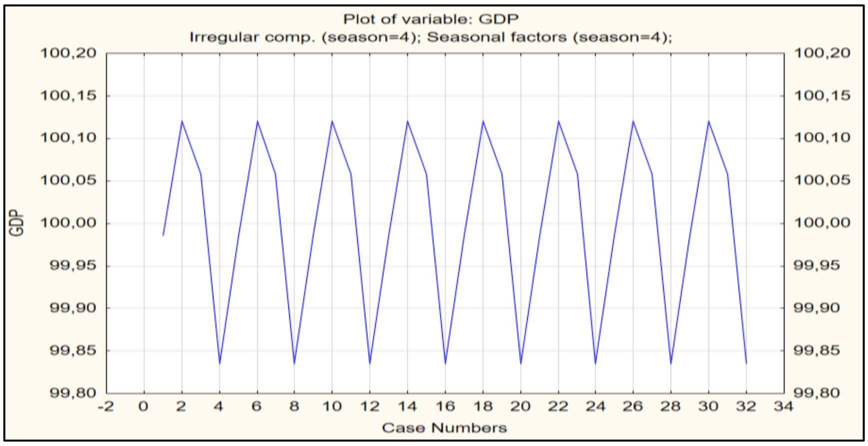


Fig. 4.13. **The seasonal component**

Then we need to build a random component graph (Fig. 4.14).

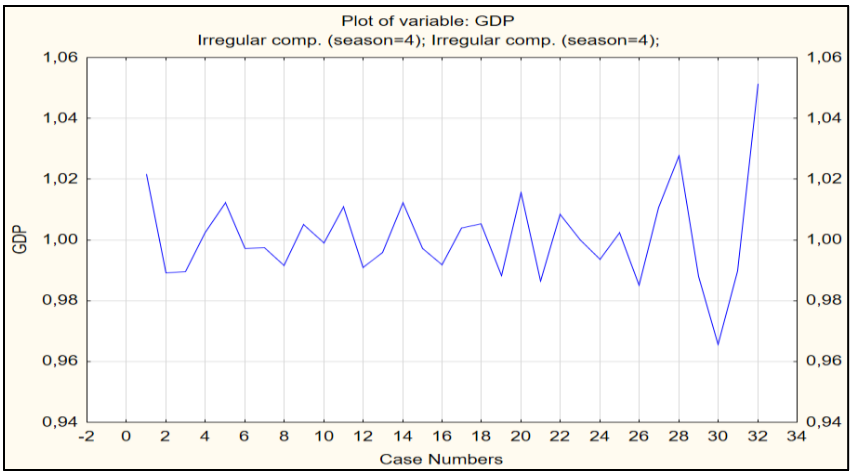


Fig. 4.14. **The random component**

The next step is to construct a regression model in which the independent variable is time (T) (Fig. 4.15).

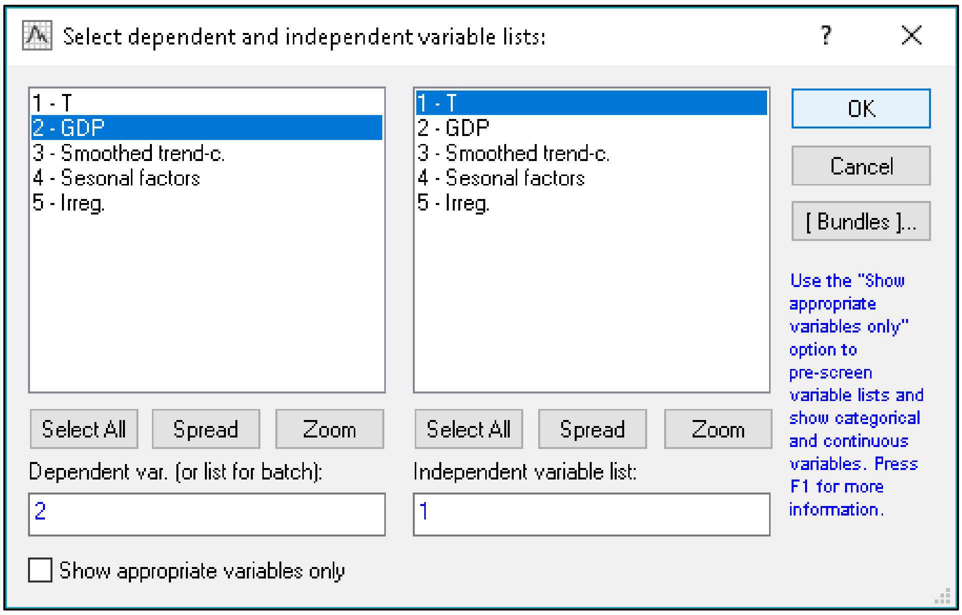


Fig. 4.15. **Choosing the parameters**

Simulation results are presented in Fig. 4.16.

Regression Summary for Dependent Variable: GDP (Spreadsheet2)
R= ,89189728 R?= ,79548076 Adjusted R?= ,78866345
F(1,30)=116,69 p<,00000 Std.Error of estimate: 59155,

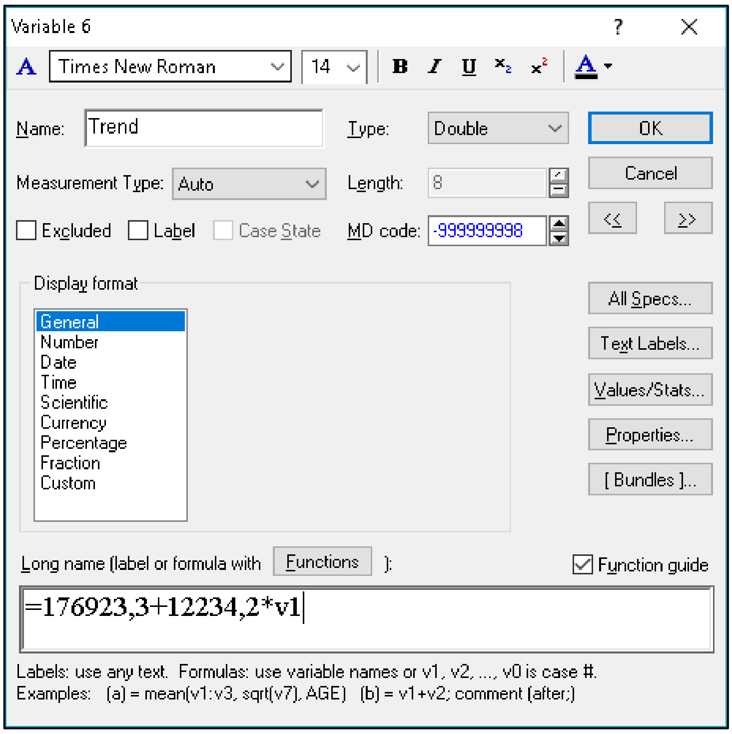
N=32	b*	Std.Err. of b*	b	Std.Err. of b	t(30)	p-value
Intercept			176923,3	21414,33	8,26191	0,000000
T	0,891897	0,082567	12234,2	1132,57	10,80210	0,000000

Fig. 4.16. **Regression model results**

So, the model will look like:

$$У = 176923.3 + 12234.2T.$$

Then, it is necessary to isolate the trend from the trend-cycle component. To do this, add a new variable and a formula for calculating it (Fig. 4.17).



	1 T	2 GDP	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 Trend
1	1	189028	219614,0	84,7703	1,015366	189157,5
2	2	214103	223577,9	97,2987	0,984208	201391,7
3	3	250306	231505,8	109,9289	0,983552	213625,9
4	4	259908	241353,6	108,0021	0,997089	225860,1
5	5	217074	252867,6	84,7703	1,012677	238094,3
6	6	255545	263435,9	97,2987	0,996977	250328,5
7	7	300446	274834,0	109,9289	0,994452	262562,7
8	8	306281	287668,0	108,0021	0,985817	274796,9
9	9	258591	303180,1	84,7703	1,006165	287031,1
10	10	310277	317393,1	97,2987	1,004720	299265,3
11	11	368488	329424,9	109,9289	1,017548	311499,5
12	12	362635	337988,8	108,0021	0,993426	323733,7
13	13	292324	344959,3	84,7703	0,999662	335967,9
14	14	346005	349745,3	97,2987	1,016771	348202,1
15	15	387109	352565,6	109,9289	0,998807	360436,3
16	16	379231	354957,1	108,0021	0,989227	372670,5
17	17	303753	357863,0	84,7703	1,001291	384904,7
18	18	354814	362692,9	97,2987	1,005436	397138,9
19	19	398000	367151,1	109,9289	0,986112	409373,1
20	20	408631	373835,5	108,0021	1,012089	421607,3
21	21	316905	380776,3	84,7703	0,981783	433841,5
22	22	382391	391161,4	97,2987	1,004719	446075,7
23	23	440476	403721,9	109,9289	0,992494	458309,9
24	24	447143	421434,1	108,0021	0,982391	470544,1
25	25	375991	445990,7	84,7703	0,994508	482778,3
26	26	456715	476229,1	97,2987	0,985649	495012,5
27	27	566997	506400,8	109,9289	1,018531	507246,7
28	28	588841	529086,8	108,0021	1,030479	519480,9
29	29	455637	547533,4	84,7703	0,981668	531715,1
30	30	535324	573077,9	97,2987	0,960054	543949,3
31	31	669170	609464,8	109,9289	0,998794	556183,5
32	32	723051	627658,3	108,0021	1,066629	568417,7
33						

Fig. 4.17. **Adding the trend component**

It is necessary to build (visualize) a trend component (Fig. 4.18).

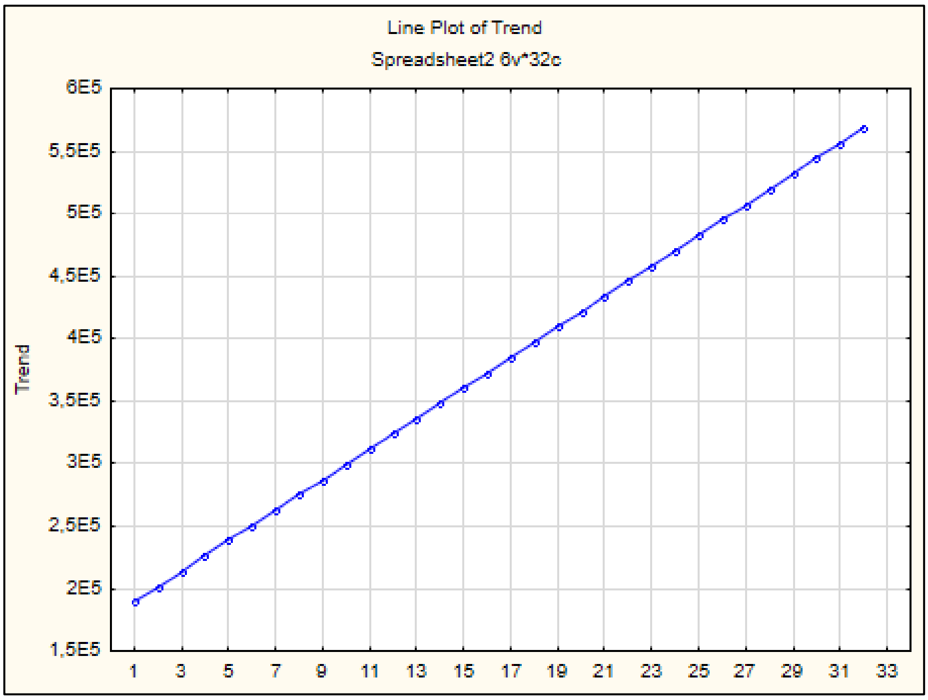


Fig. 4.18. **Visualization of the trend component**

The values of the cycle component are then calculated as follows: Cycle = Smoothed Trend-c. / Trend (Fig. 4.19).

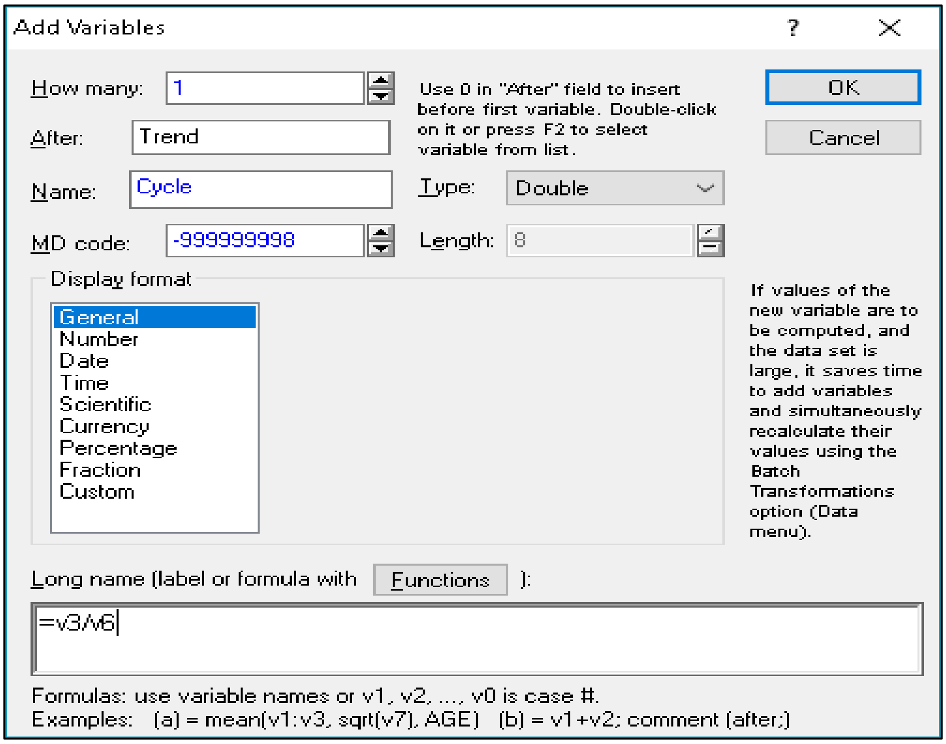


Fig. 4.19. **Adding the cycle component**

The window for entering a new variable is shown in Fig. 4.20.

	1 T	2 GDP	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 Trend	7 Cycle
1	1	189028	219614,0	84,7703	1,015366	189157,5	1,161011
2	2	214103	223577,9	97,2987	0,984208	201391,7	1,110164
3	3	250306	231505,8	109,9289	0,983552	213625,9	1,083697
4	4	259908	241353,6	108,0021	0,997089	225860,1	1,068598
5	5	217074	252867,6	84,7703	1,012677	238094,3	1,062048
6	6	255545	263435,9	97,2987	0,996977	250328,5	1,052361
7	7	300446	274834,0	109,9289	0,994452	262562,7	1,046737
8	8	306281	287668,0	108,0021	0,985817	274796,9	1,046839
9	9	258591	303180,1	84,7703	1,006165	287031,1	1,056262
10	10	310277	317393,1	97,2987	1,004720	299265,3	1,060574
11	11	368488	329424,9	109,9289	1,017548	311499,5	1,057546
12	12	362635	337988,8	108,0021	0,993426	323733,7	1,044033
13	13	292324	344959,3	84,7703	0,999662	335967,9	1,026763
14	14	346005	349745,3	97,2987	1,016771	348202,1	1,004432
15	15	387109	352565,6	109,9289	0,998807	360436,3	0,978163
16	16	379231	354957,1	108,0021	0,989227	372670,5	0,952469
17	17	303753	357863,0	84,7703	1,001291	384904,7	0,929744
18	18	354814	362692,9	97,2987	1,005436	397138,9	0,913264
19	19	398000	367151,1	109,9289	0,986112	409373,1	0,896862
20	20	408631	373835,5	108,0021	1,012089	421607,3	0,886691
21	21	316905	380776,3	84,7703	0,981783	433841,5	0,877685
22	22	382391	391161,4	97,2987	1,004719	446075,7	0,876895
23	23	440476	403721,9	109,9289	0,992494	458309,9	0,880893
24	24	447143	421434,1	108,0021	0,982391	470544,1	0,895632
25	25	375991	445990,7	84,7703	0,994508	482778,3	0,9238
26	26	456715	476229,1	97,2987	0,985649	495012,5	0,962055
27	27	566997	506400,8	109,9289	1,018531	507246,7	0,998332
28	28	588841	529086,8	108,0021	1,030479	519480,9	1,018491
29	29	455637	547533,4	84,7703	0,981668	531715,1	1,02975
30	30	535324	573077,9	97,2987	0,960054	543949,3	1,05355
31	31	669170	609464,8	109,9289	0,998794	556183,5	1,095798
32	32	723051	627658,3	108,0021	1,066629	568417,7	1,10422
33							

Fig. 4.20. **The results of adding the cycle component**

The graph of the cyclic component is presented in Fig. 4.21.

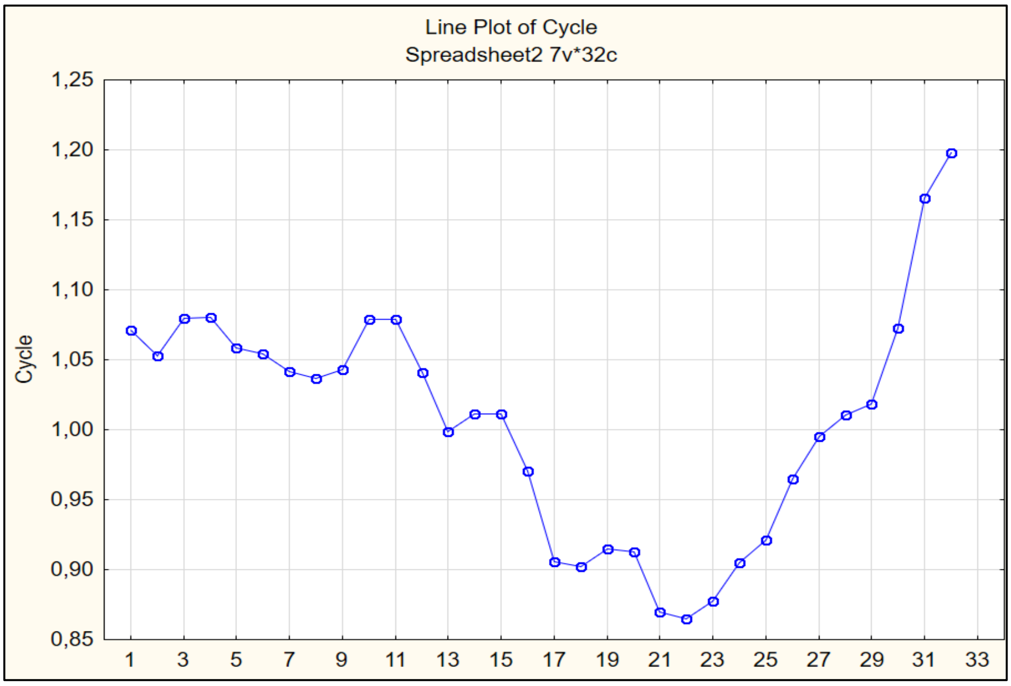


Fig. 4.21. **Visualization of the cycle component**

Before proceeding to forecast GDP for 4 periods ahead using the time series decomposition model, it is necessary to perform a number of actions:

- add 4 observations after the last available in the series;
- in the data column T (time period), enter the corresponding ordinal numbers, continuing the series;
- in the column *Seasonal Factors*, enter the corresponding values of the seasonal components;
- in the *Cycle* column, enter the corresponding cyclic component values, taking into account the cycle period;
- in the *Trend* column set the data recalculation (Fig. 4.22).

	1 T	2 GDP	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 Trend	7 Cycle
1	1	189028	219614,0	84,7703	1,015366	189157,5	1,161011
2	2	214103	223577,9	97,2987	0,984208	201391,7	1,110164
3	3	250306	231505,8	109,9289	0,983552	213625,9	1,083697
4	4	259908	241353,6	108,0021	0,997089	225860,1	1,068598
5	5	217074	252867,6	84,7703	1,012677	238094,3	1,062048
6	6	255545	263435,9	97,2987	0,996977	250328,5	1,052361
7	7	300446	274834,0	109,9289	0,994452	262562,7	1,046737
8	8	306281	287668,0	108,0021	0,985817	274796,9	1,046839
9	9	258591	303180,1	84,7703	1,006165	287031,1	1,056262
10	10	310277	317393,1	97,2987	1,004720	299265,3	1,060574
11	11	368488	329424,9	109,9289	1,017548	311499,5	1,057546
12	12	362635	337988,8	108,0021	0,993426	323733,7	1,044033
13	13	292324	344959,3	84,7703	0,999662	335967,9	1,026763
14	14	346005	349745,3	97,2987	1,016771	348202,1	1,004432
15	15	387109	352565,6	109,9289	0,998807	360436,3	0,978163
16	16	379231	354957,1	108,0021	0,989227	372670,5	0,952469
17	17	303753	357863,0	84,7703	1,001291	384904,7	0,929744
18	18	354814	362692,9	97,2987	1,005436	397138,9	0,913264
19	19	398000	367151,1	109,9289	0,986112	409373,1	0,896862
20	20	408631	373835,5	108,0021	1,012089	421607,3	0,886691
21	21	316905	380776,3	84,7703	0,981783	433841,5	0,877685
22	22	382391	391161,4	97,2987	1,004719	446075,7	0,876895
23	23	440476	403721,9	109,9289	0,992494	458309,9	0,880893
24	24	447143	421434,1	108,0021	0,982391	470544,1	0,895632
25	25	375991	445990,7	84,7703	0,994508	482778,3	0,9238
26	26	456715	476229,1	97,2987	0,985649	495012,5	0,962055
27	27	566997	506400,8	109,9289	1,018531	507246,7	0,998332
28	28	588841	529086,8	108,0021	1,030479	519480,9	1,018491
29	29	455637	547533,4	84,7703	0,981668	531715,1	1,02975
30	30	535324	573077,9	97,2987	0,960054	543949,3	1,05355
31	31	669170	609464,8	109,9289	0,998794	556183,5	1,095798
32	32	723051	627658,3	108,0021	1,066629	568417,7	1,10422
33	33			84,7703		580651,9	1,060574
34	34			97,2987		592886,1	1,057546
35	35			109,9289		605120,3	1,044033
36	36			108,0021		617354,5	1,026763
37							

Fig. 4.22. **The steps of the analysis**

Then we need to add a new variable of the GDP prediction (Fig. 4.23, 4.24).

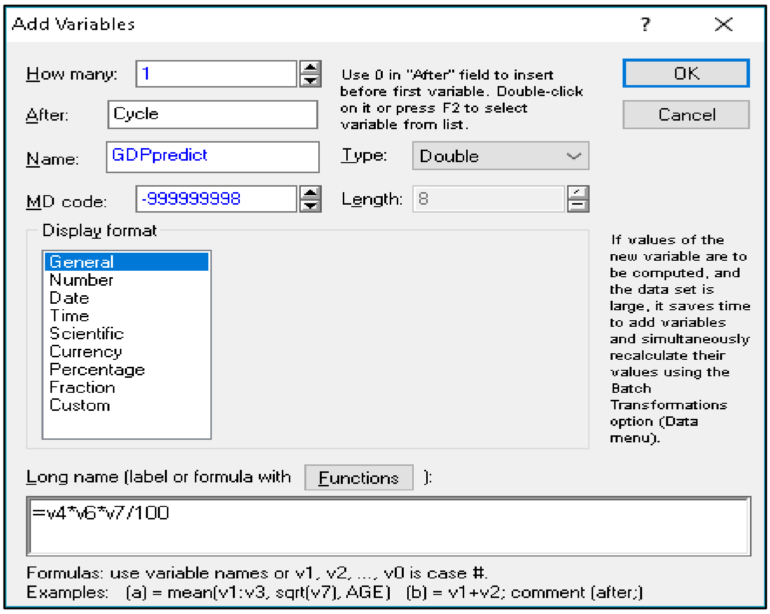


Fig. 4.23. **Adding the GDP prediction**

Then you can calculate the forecast values of the GDP indicator by 5 steps forward by specifying a model of the form:

$$\text{GDPpredict} = \text{Trend} \times \text{Cycle} \times \text{Seasonal Factors} / 100.$$

	1 T	2 GDP	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 Trend	7 Cycle	8 GDP predict
1	1	189028	219614,0	84,7703	1,015366	189157,5	1,161011	186167,3
2	2	214103	223577,9	97,2987	0,984208	201391,7	1,110164	217538,4
3	3	250306	231505,8	109,9289	0,983552	213625,9	1,083697	254491,8
4	4	259908	241353,6	108,0021	0,997089	225860,1	1,068598	260666,9
5	5	217074	252867,6	84,7703	1,012677	238094,3	1,062048	214356,5
6	6	255545	263435,9	97,2987	0,996977	250328,5	1,052361	256319,8
7	7	300446	274834,0	109,9289	0,994452	262562,7	1,046737	302122
8	8	306281	287668,0	108,0021	0,985817	274796,9	1,046839	310687,4
9	9	258591	303180,1	84,7703	1,006165	287031,1	1,056262	257006,6
10	10	310277	317393,1	97,2987	1,004720	299265,3	1,060574	308819,4
11	11	368488	329424,9	109,9289	1,017548	311499,5	1,057546	362133,3
12	12	362635	337988,8	108,0021	0,993426	323733,7	1,044033	365034,9
13	13	292324	344959,3	84,7703	0,999662	335967,9	1,026763	292422,9
14	14	346005	349745,3	97,2987	1,016771	348202,1	1,004432	340297,7
15	15	387109	352565,6	109,9289	0,998807	360436,3	0,978163	387571,6
16	16	379231	354957,1	108,0021	0,989227	372670,5	0,952469	383361,1
17	17	303753	357863,0	84,7703	1,001291	384904,7	0,929744	303361,4
18	18	354814	362692,9	97,2987	1,005436	397138,9	0,913264	352895,5
19	19	398000	367151,1	109,9289	0,986112	409373,1	0,896862	403605,3
20	20	408631	373835,5	108,0021	1,012089	421607,3	0,886691	403750,2
21	21	316905	380776,3	84,7703	0,981783	433841,5	0,877685	322785,1
22	22	382391	391161,4	97,2987	1,004719	446075,7	0,876895	380595
23	23	440476	403721,9	109,9289	0,992494	458309,9	0,880893	443807,1
24	24	447143	421434,1	108,0021	0,982391	470544,1	0,895632	455157,6
25	25	375991	445990,7	84,7703	0,994508	482778,3	0,9238	378067,5
26	26	456715	476229,1	97,2987	0,985649	495012,5	0,962055	463364,9
27	27	566997	506400,8	109,9289	1,018531	507246,7	0,998332	556681
28	28	588841	529086,8	108,0021	1,030479	519480,9	1,018491	571424,8
29	29	455637	547533,4	84,7703	0,981668	531715,1	1,02975	464145,5
30	30	535324	573077,9	97,2987	0,960054	543949,3	1,05355	557597,5
31	31	669170	609464,8	109,9289	0,998794	556183,5	1,095798	669978,2
32	32	723051	627658,3	108,0021	1,066629	568417,7	1,10422	677884
33	33			84,7703		580651,9	1,060574	522036
34	34			97,2987		592886,1	1,057546	610067
35	35			109,9289		605120,3	1,044033	694493,3
36	36			108,0021		617354,5	1,026763	684599,9

Fig. 4.24. **The results of adding the GDP prediction**

Let's build a graph of the GDP prediction (Fig. 4.25).

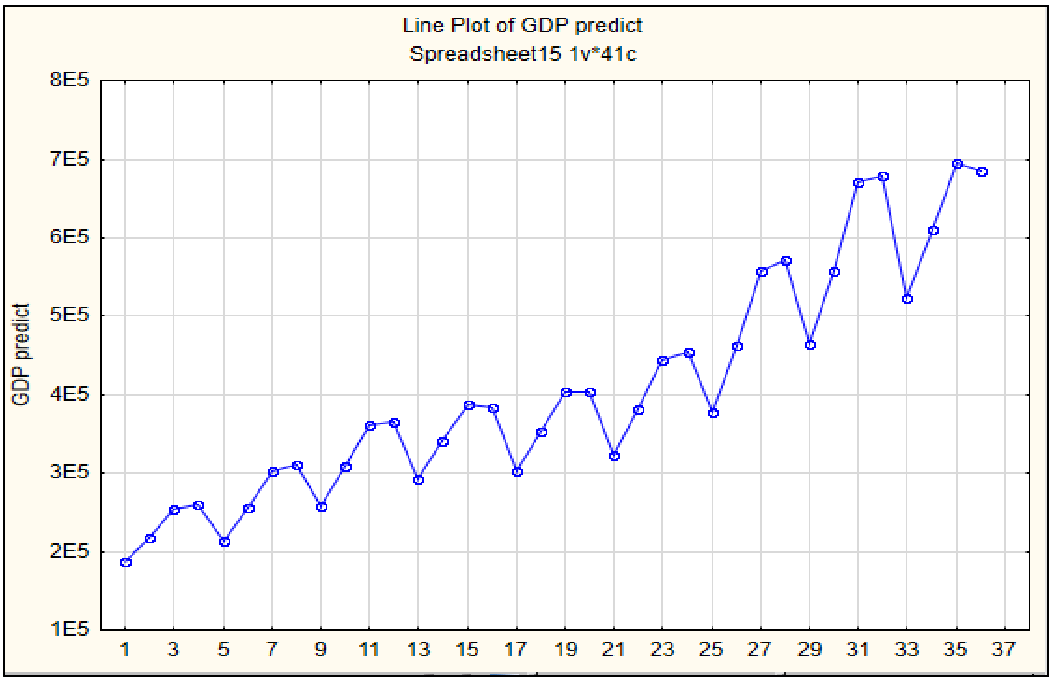


Fig. 4.25. **Visualization of the GDP prediction**

Fig. 4.26 shows the histogram of error distribution. The fact that this distribution is close to the normal law is a confirmation of the adequacy of the model and the accuracy of the forecast.

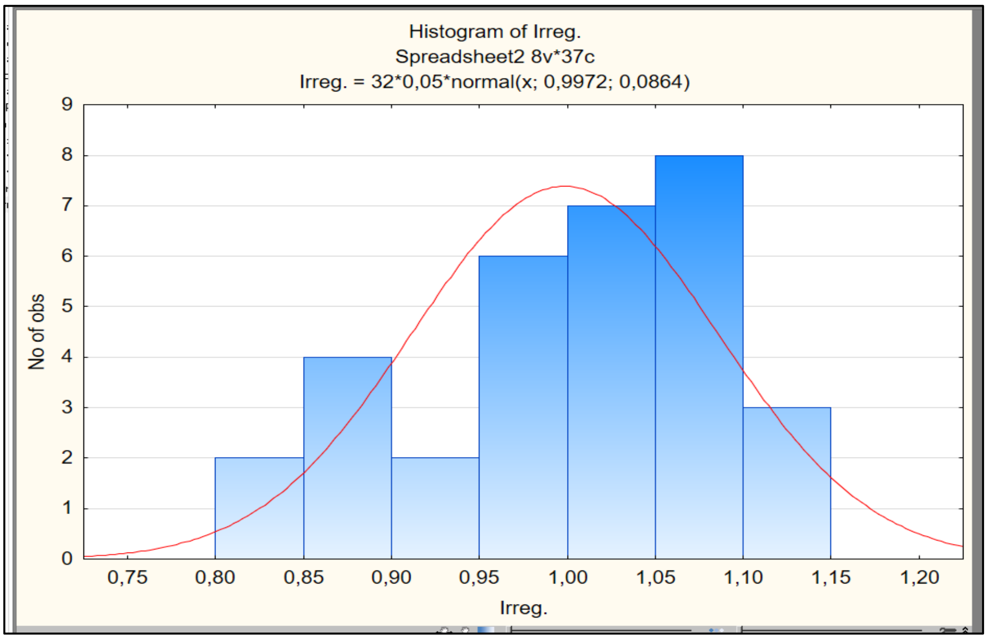


Fig. 4.26. **The histogram of the GDP prediction with the curve of normal distribution**

To confirm the quality of the forecast, we calculate the average relative percentage error by formula (4.2).

$$\text{MAPE} = \frac{1}{n} \times \sum \frac{|y - \bar{y}|}{y} \times 100\ \%, \qquad (4.2)$$

where MAPE is a measure of the bias of the forecast (forecasting errors of the time series).

If MAPE < 10 %, this will indicate a high accuracy of the made forecast.

Calculations can be performed in Excel (Fig. 2.27).

E36 =(E35*100)/32

	A	B	C	D	E
1					
2	GDP	GDP predict	y-y pr	ABS	(y - y pr)/y
3	189028	186167,3261	2860,67	2860,67	0,0151336
4	214103	217538,4377	-3435,44	3435,44	0,0160457
5	250306	254491,8123	-4185,81	4185,81	0,0167228
6	259908	260666,8713	-758,871	758,871	0,0029198
7	217074	214356,519	2717,48	2717,48	0,0125187
8	255545	256319,7642	-774,764	774,764	0,0030318
9	300446	302122,0424	-1676,04	1676,04	0,0055785
10	306281	310687,415	-4406,41	4406,41	0,0143868
11	258591	257006,5703	1584,43	1584,43	0,0061272
12	310277	308819,4313	1457,57	1457,57	0,0046976
13	368488	362133,303	6354,7	6354,7	0,0172453
14	362635	365034,8861	-2399,89	2399,89	0,0066179
15	292324	292422,9127	-98,9127	98,9127	0,0003384
16	346005	340297,7132	5707,29	5707,29	0,0164948
17	387109	387571,552	-462,552	462,552	0,0011949
18	379231	383361,0963	-4130,1	4130,1	0,0108907
19	303753	303361,4101	391,59	391,59	0,0012892
20	354814	352895,5216	1918,48	1918,48	0,005407
21	398000	403605,2537	-5605,25	5605,25	0,0140836
22	408631	403750,1655	4880,83	4880,83	0,0119444
23	316905	322785,1091	-5880,11	5880,11	0,0185548
24	382391	380595,0498	1795,95	1795,95	0,0046966
25	440476	443807,1305	-3331,13	3331,13	0,0075626
26	447143	455157,6489	-8014,65	8014,65	0,0179241
27	375991	378067,4679	-2076,47	2076,47	0,0055227
28	456715	463364,8874	-6649,89	6649,89	0,0145603
29	566997	556680,9723	10316	10316	0,0181941
30	588841	571424,7623	17416,2	17416,2	0,0295771
31	455637	464145,5336	-8508,53	8508,53	0,0186739
32	535324	557597,5209	-22273,5	22273,5	0,0416076
33	669170	669978,1514	-808,151	808,151	0,0012077
34	723051	677883,9919	45167	45167	0,0624673
35				Sum	0,4232174
36				MAPE	1,3225543

Fig. 2.27. **Calculations of the MAPE**

The average relative percentage error is 1.323 %, which is confirmed by the accuracy of the forecast for this multiplicative time series model.

As an output, it is necessary to give an economic interpretation of the results of forecasting, i.e. say what will happen to the country's GDP in the near future.

Content module 2. Modeling and forecasting of multidimensional processes

Topic 5. Factor analysis of data

Laboratory work 5
Building a model of factor analysis

The purpose of the work is to acquire data processing skills using the factor analysis methods.

The task is to reduce the information space using the methods of factor analysis.

Guidelines

The *Factor Analysis* module contains a wide range of methods that allow the selection of factors, thus reducing the input information space.

Let's consider the main stages of conducting factor analysis in the system (package) Statistica based on the following example. For the analysis of the activity of a private enterprise, the following indicators were selected (Table 5.1): X1, the rate of the losses due to spoilage; X2, the index of reduction of production cost; X3, the return on capital; X4, the coefficient of the equipment variability; X5, productivity; X6, the share of the purchased items.

Table 5.1

The initial data

Period	X1	X2	X3	X4	X5	X6
1	2	3	4	5	6	7
1	5.571	47.88	0.522	0.153	0.071	0.225
2	4.914	27.09	1.377	0.135	0.873	0.441
3	5.85	131.76	0.63	0.144	1.035	0.234
4	5.949	16.29	1.593	0.135	0.018	0.252
5	3.888	12.24	0.666	0.153	0.054	0.153
6	6.633	80.82	0.972	0.306	1.251	0.153
7	6.318	56.25	1.035	0.306	0.072	0.279

Table 5.1 (the end)

1	2	3	4	5	6	7
8	7.425	41.67	0.873	0.306	0.693	0.162
9	7.335	93.123	1.008	0.171	0.693	0.279
10	7.848	65.97	0.891	0.171	0.972	0.162
11	5.976	68.94	0.522	0.306	0.837	0.279
12	7.29	65.709	0.927	0.306	0.09	0.135
13	4.968	29.07	1.116	0.135	0.099	0.252
14	8.433	178.686	0.801	0.171	1.296	0.162
15	11.853	538.308	0.612	0.306	0.432	0.126
16	6.003	64.521	0.927	0.171	0.116	0.162
17	5.112	81.567	0.657	0.288	0.693	0.261
18	4.671	73.89	0.657	0.171	0.837	0.27
19	9.018	68.58	0.765	0.297	0.117	0.243
20	7.344	107.523	0.927	0.306	1.557	0.261
21	4.671	73.89	0.657	0.171	0.837	0.27

To reduce the initial information space we need to use *Statistics / Multivariate Exploratory Techniques / Factor Ananlysis (Multidimensional Methods / Factor Analysis)* to select the *Factor Analysis* module. The *Factor Analysis* dialog box appears (Fig. 5.1).

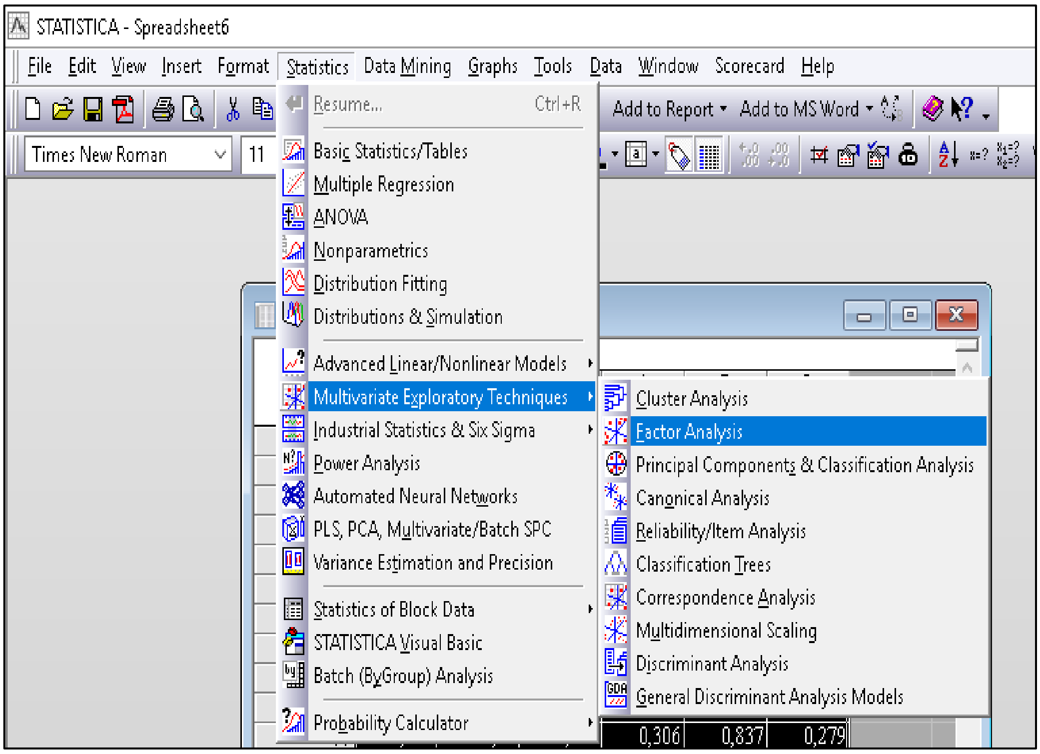


Fig. 5.1. **The *Factor Analysis* module**

The *Variables* button allows you to select all variables from the data file that must be included in the factor analysis. If all variables are used for analysis, you can use the *Select All* button (Fig. 5.2).

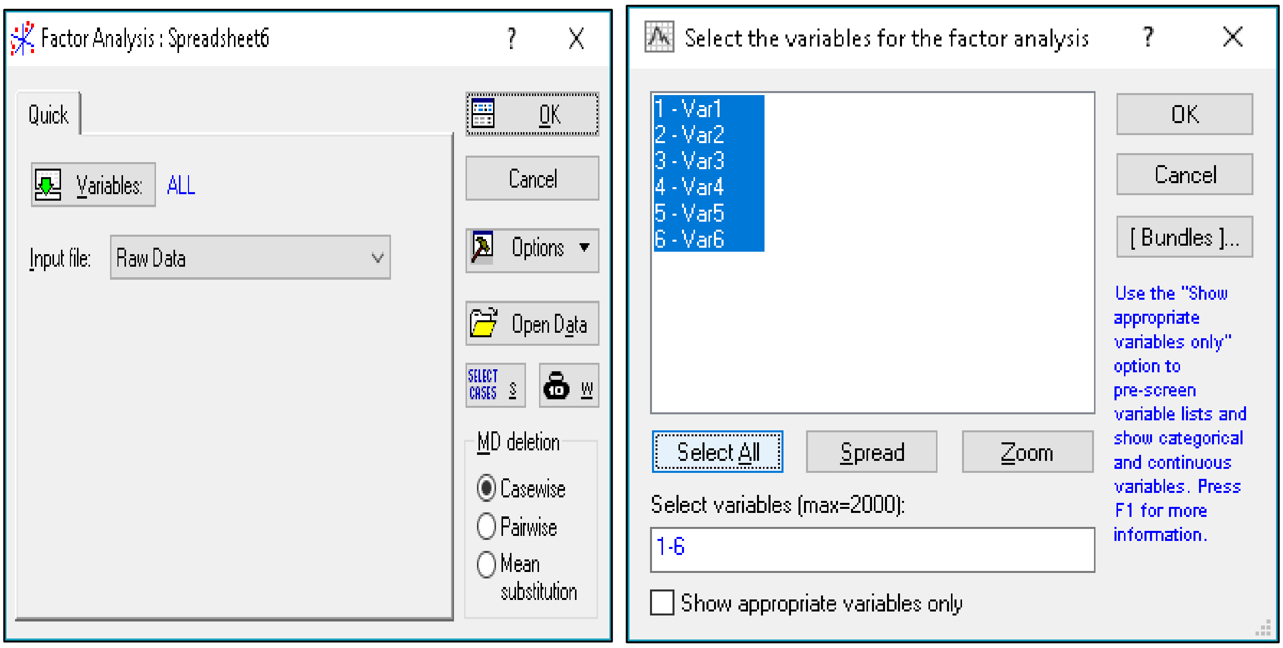


Fig. 5.2. **Choosing the variables**

The module includes the following output types: *Correlation Matrix* and *Raw Data*.

Choose, for example, *Raw Data*. This is a regular data file, where the values of the variables are given in rows.

MD deletion (replace missed variables) is used to choose the method for processing the missed values.

The *Casewise* option (the way to exclude the missed cases) implies that in the spreadsheet containing the data all the rows (cases) that have at least one missing value are ignored. This applies to all variables. In the table, there are only cases in which there is no skip.

The *Pairwise* option (a duplicate way to exclude the missed values). Missed cases are ignored for all variables, but only for the selected pair. All cases in which there are no spaces are used in processing, for example, with elemental calculation of the correlation matrix, when all pairs of variables are sequentially considered. Obviously, the pairwise method has more observations for processing than the *Casewise* way.

Mean Substitution is substitution of the average instead of the missed values.

By clicking on the OK button in the startup window of the module, the analysis of the selected variables begins. The Statistica system will process the missed values in the way indicated, will calculate the correlation matrix and will offer a choice of several methods for factor analysis. The calculation of the correlation matrix (if not specified immediately) is the first stage of factor analysis. After clicking the Ok button, you can go to the next dialog.

The window *Define Method of Factor Extraction* (determining the method of selection of factors) is presented in Fig. 5.3.

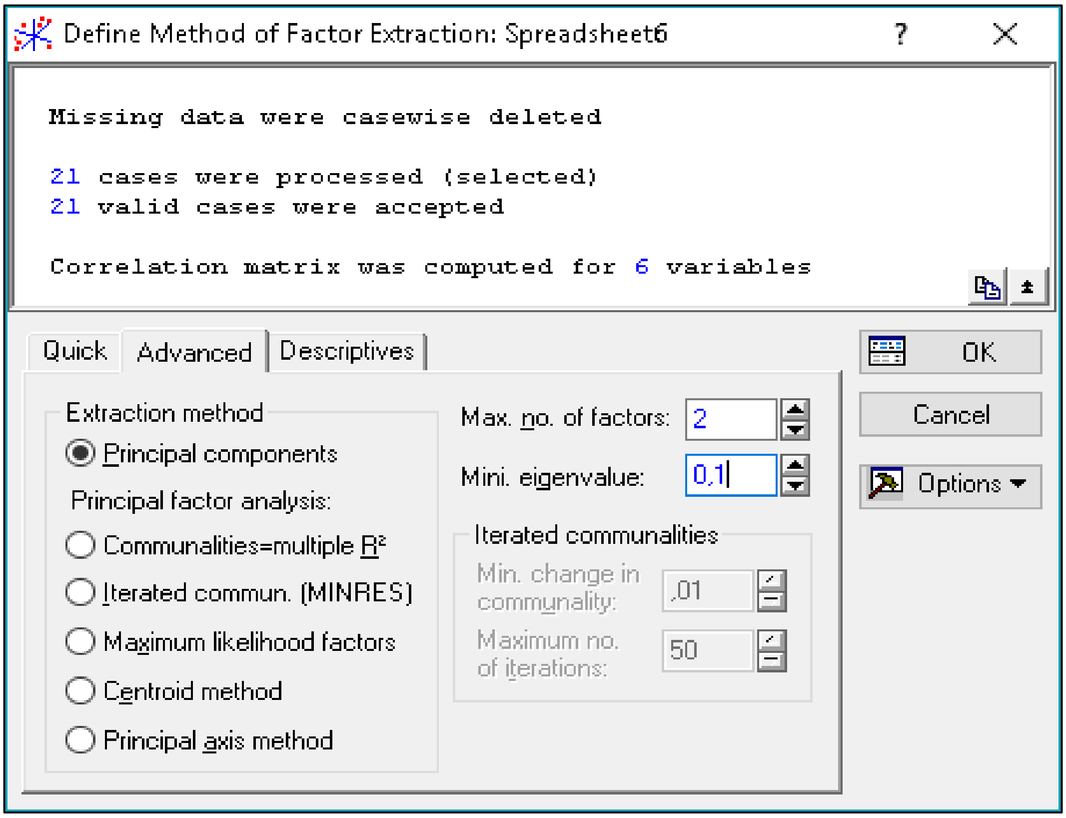


Fig. 5.3. **Defining the method of factor extraction**

This window has the following structure. The upper part of the window is informational: it is reported here that the missing values are processed by the casewise method. 21 cases were processed and 21 cases were taken for further calculations. The correlation matrix was calculated for 6 variables. The group of options merged under the heading *Extraction method* allows you to choose the method of processing.

To continue the analysis in the *Define Method of Factor Extraction window* (Fig. 5.4), we need to click on the *Review correlations, means, standard deviations* button.

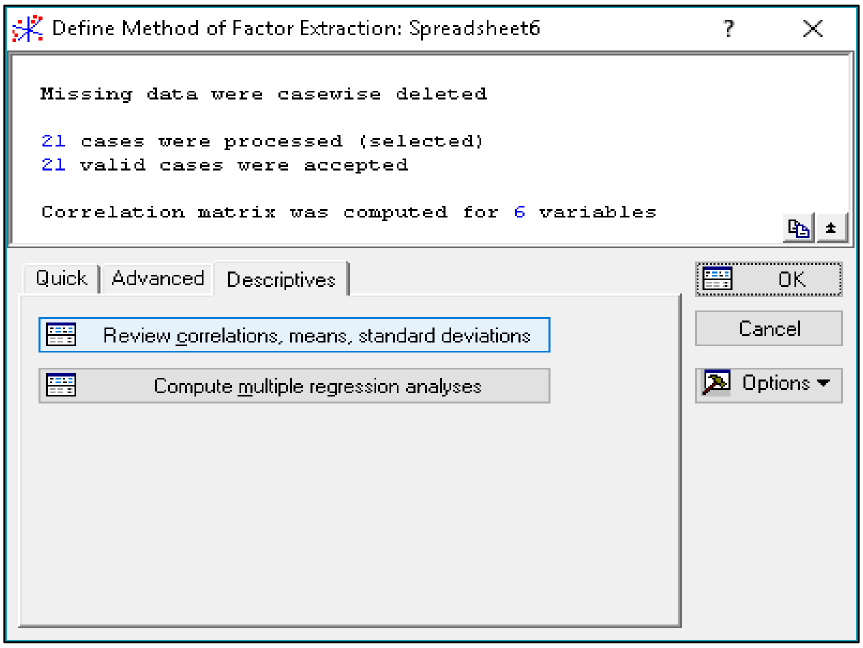


Fig. 5.4. **Review of correlations, means, standard deviations**

After that, a window for viewing descriptive statistics for the analyzed data appears, where you can see the average, standard deviations, correlations, covariations, build different graphs (Fig. 5.5).

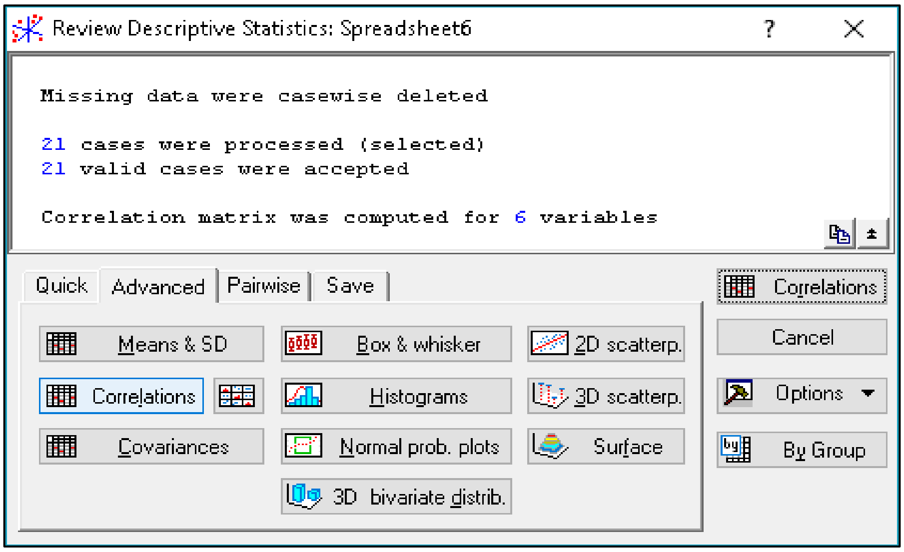


Fig. 5.5. **Additional analysis**

Here one can carry out additional analysis of the current data, verify the conformity of the sample variables to the normal distribution law and the existence of a linear correlation between the variables. Clicking the *Correlations* button will display the correlation matrix of the variables selected earlier (Fig. 5.6).

Variable	Correlations (Spreadsheet6) Casewise deletion of MD N=21					
Variable	**Var1**	Var2	Var3	Var4	Var5	Var6
Var1	1,00	0,75	-0,09	0,49	0,11	-0,45
Var2	0,75	1,00	-0,32	0,29	0,15	-0,35
Var3	-0,09	-0,32	1,00	-0,24	-0,14	0,33
Var4	0,49	0,29	-0,24	1,00	0,12	-0,26
Var5	0,11	0,15	-0,14	0,12	1,00	0,12
Var6	-0,45	-0,35	0,33	-0,26	0,12	1,00

Fig. 5.6. **The correlation matrix**

So, at the next stage, we choose the method of specification of factors – the method of the principal components (taking as a basis the correlation matrix of the initial data, this method helps to reduce the dimension of the data and minimize the loss of information) – and specify the maximum number of factors (in our case it is 6) and the minimum actual value of the Kaiser criterion, to be not less than 1. Statistica 8.0 automatically performs calculations and displays the results of the factor analysis (Fig. 5.7).

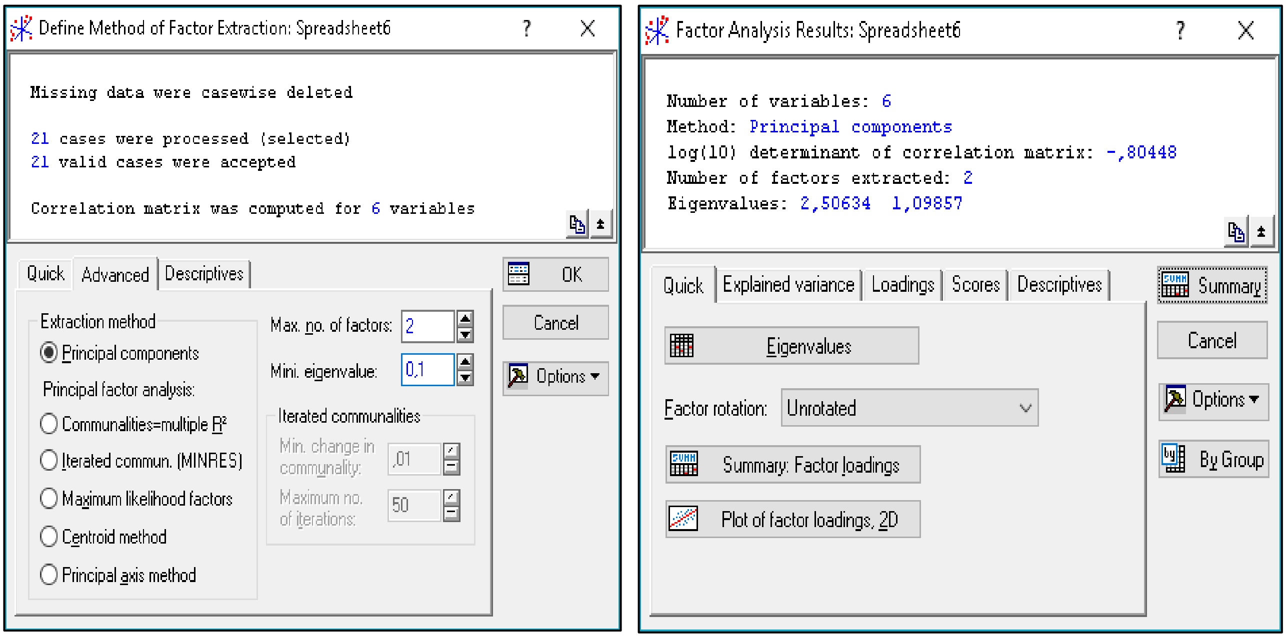


Fig. 5.7. **Defining the methods of factor analysis**

In the upper part of the window of the results of factor analysis, an informative message is given: *Number of variables* (the number of the analyzed variables) – 6; *Method* (the method of analysis) – the principal components; *log (10) determination of the correlation matrix* (a decimal logarithm of the determinant of the correlation matrix) – 0.80448; *Number of factors extracted* (the number of the selected factors) – 2; *Eigenvalues* – 2.50634; 1.09857.

As a result, two main factors that correspond to the highest eigenvalues of the correlation matrix were identified: λ1 = 2.50634 and λ2 = 1.09857, so two of these factors account for the largest part (60.08 %) of the variance explanation. Namely, the first factor explains approximately 42 % (41.77 %) of the total dispersion, while the share of the second factor accounts for almost 18 % (18.31 %) of the dispersion explanation. Together, they describe approximately 60 % of the dispersion, that is, almost the entire array of data (Fig. 5.8).

	Eigenvalues (Spreadsheet6) Extraction: Principal components			
Value	**Eigenvalue**	% Total variance	Cumulative Eigenvalue	Cumulative %
1	2,506340	41,77234	2,506340	41,77234
2	1,098566	18,30944	3,604907	60,08178

Fig. 5.8. **The eigenvalues**

Thus, factorization is almost complete, although there are other, less significant factors. In order to ensure that the correct number of factors is obtained, it is expedient to use the Kettel criterion (stony maturity criterion), which makes it possible to reflect graphically the eigenvalues of each selected factor in descending order and find on the graph a place where the reduction of these values from the left to the right as much as possible slows down. In accordance with this criterion, at the points with the coordinates 1, 2 the stony maturity is slowed down most significantly, therefore, theoretically, one can restrict two factors (Fig. 5.9 and 5.10).

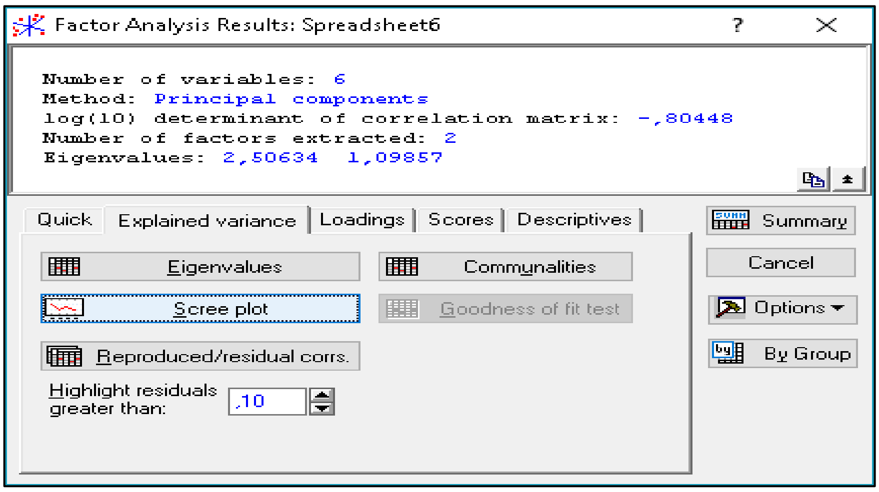


Fig. 5.9. **Choosing the visualization of eigenvalues**

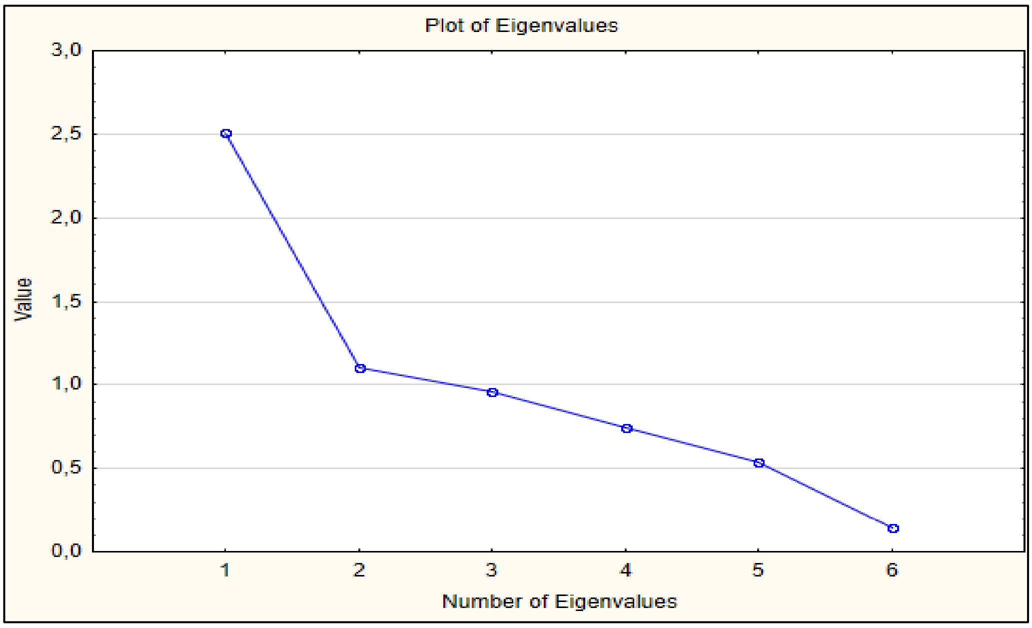


Fig. 5.10. **The plot of eigenvalues**

At the bottom of the window there are subdivisions that allow you to comprehensively get acquainted with the results of the analysis numerically and graphically. The options *Plot of loadings, 2D and Plot of loadings, 3D (Load Charts)* will plot factor load graphs in the projection onto the plane of any two selected factors and in the projection onto the space of the three selected factors (Fig. 5.11).

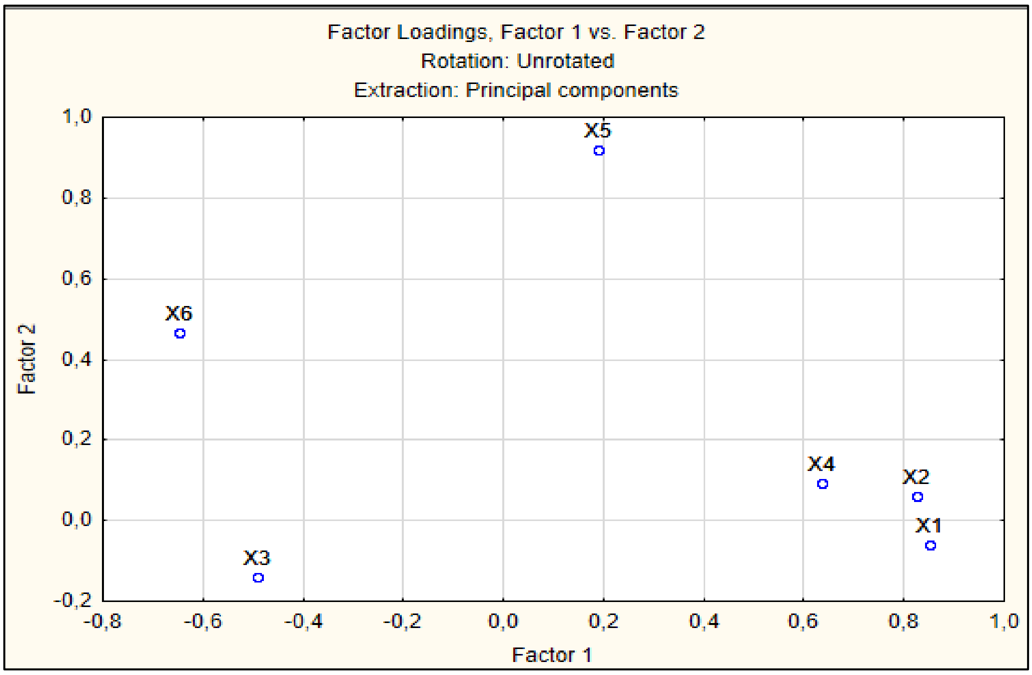


Fig. 5.11. **The polygon of factor loadings**

Summary. The option *Factor loadings* makes up a table with current factor loads that are calculated for this method of rotation of factors. In this table, the factors correspond to the columns, and the variables are rows, and for each factor, the load of each output variable is given, which shows the relative magnitude of the projection of the variable on the factor coordinate axis. Factor loads can be interpreted as correlations between the corresponding variables and factors – the higher the load modulus, the greater the proximity of the factor to the initial variable; and they represent the most important information for interpreting the resulting factors. In a generated table, for facilitation purposes, factor loadings will be presented in the absolute value greater than 0.7 (Fig. 5.12).

	Factor Loadings (Unrotated) (Spreadsheet6) Extraction: Principal components (Marked loadings are >,700000)	
Variable	**Factor 1**	Factor 2
Var1	0,853123	-0,059150
Var2	0,824095	0,059803
Var3	-0,490487	-0,141242
Var4	0,636260	0,093964
Var5	0,189549	0,918940
Var6	-0,646573	0,467185
Expl.Var	2,506340	1,098566
Prp.Totl	0,417723	0,183094

Fig. 5.12**. Factor loadings**

The results presented in Fig. 5.12 show that the first factor is more correlated with variables than the second one. Since the correlation of other factors is insignificant, it is advisable to resort to the rotation of the axes, hoping to obtain a solution that can be interpreted in the subject area. Now it is expedient to determine which indicators were included in the first and second factors. To do this, you should review the factor loads (correlations between the corresponding variables and factors – the higher the load modulus, the greater the proximity of the factor to the output variable). However, one should immediately turn to the axes in order to obtain a simple structure in which most observations are located near the axes of coordinates.

The results of changes in the composition of factors and factor loads after their rotation using the normalized version are given in Fig. 5.13.

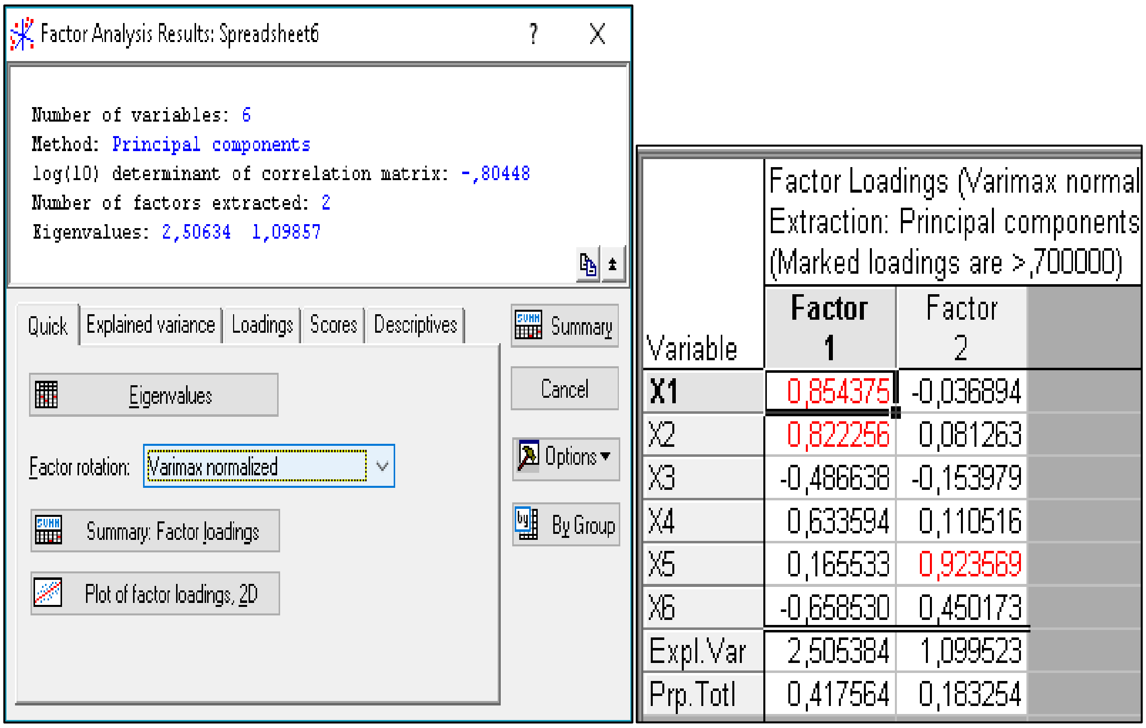


Fig. 5.13. **The results of changes in the composition of factors and factor loads after their rotation**

Thus, according to the results we have the following equations, where X1 – X6 are indicators characterizing the activity of a private enterprise and f1, f2 are factor loads:

X1 = 0.854 f1 + 0.309 f2,
X2 = 0.822 f1 + 0.081 f2,
X3 = 0.487 f1 + 0.154 f2,

X11 = 0.634 f1 + 0.111 f2,
X12 = 0.166 f1 + 0.924 f2,
X13 = 0.659 f1 + 0.450 f2.

The criterion applied for selecting quantitative indicators in the integral was factor loads the value of which was determined by the Pearson correlation coefficient (R) to be 0.7 ≥ R > 0.9, which, according to the Chaddock scale, indicates a strong correlation between the investigated parameters. It means that for the analyzed period, 60 % of changes in the indicators of activity of a private enterprise are explained by the influence of the following indicators: X1, X2 and X5.

Topic 6. Cluster analysis as a means of forming homogeneous data groups

Laboratory work 6
Using cluster analysis for the study of economic processes

The purpose of the work is to get skills in the use of cluster analysis in the Statistica package.

The task is to carry out the classification of countries of the world, according to the level of energy security.

Guidelines

The level of countries' energy security is assessed based on the following indicators:

1) the share of their own sources in the balance of fuel and energy resources,% (X1);

2) the share of the dominant fuel resource in the consumption of fuel and energy resources,% (X2);

3) energy intensity of GDP, kg of conditional fuel / UAH, (X3);

4) the volume of coal production, million tons (X4);

5) the degree of supply of fuel and energy resources (X5).

The output values for these indicators are shown in Table 6.1.

Table 6.1

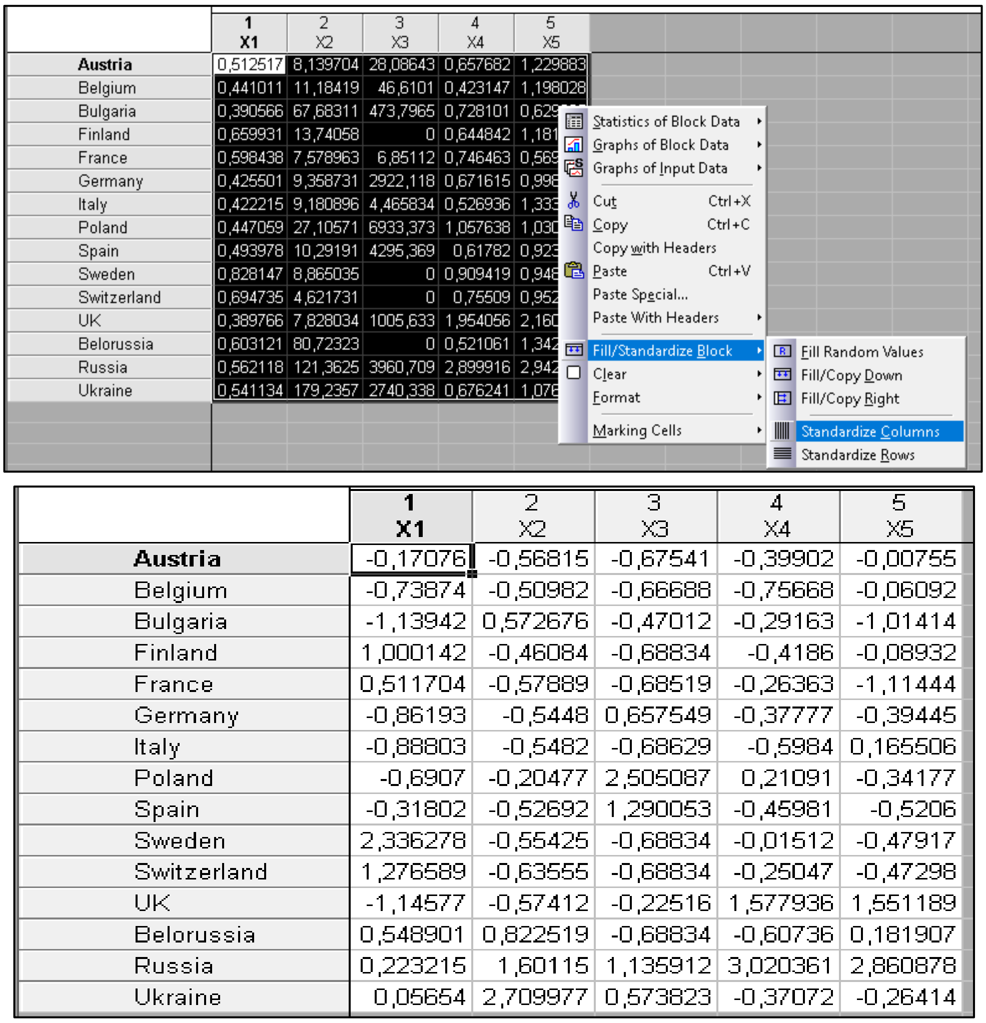
The significance of energy security indicators

Countries	Indicators of countries' energy security				
	X1	X2	X3	X4	X5
1	2	3	4	5	6
Austria	0.512517	8.139704	28.08643	0.657682	1.229883
Belgium	0.441011	11.18419	46.6101	0.423147	1.198028
Bulgaria	0.390566	67.68311	473.7965	0.728101	0.629035
Finland	0.659931	13.74058	0	0.644842	1.181073
France	0.598438	7.578963	6.85112	0.746463	0.569163
Germany	0.425501	9.358731	2922.118	0.671615	0.998937

Table 6.1 (the end)

1	2	3	4	5	6
Italy	0.422215	9.180896	4.465834	0.526936	1.333184
Poland	0.447059	27.10571	6933.373	1.057638	1.030381
Spain	0.493978	10.29191	4295.369	0.61782	0.923633
Sweden	0.828147	8.865035	0	0.909419	0.948364
Switzerland	0.694735	4.621731	0	0.75509	0.952061
The UK	0.389766	7.828034	1005.633	1.954056	2.160322
Belorussia	0.603121	80.72323	0	0.521061	1.342974
Russia	0.562118	121.3625	3960.709	2.899916	2.942098
Ukraine	0.541134	179.2357	2740.338	0.676241	1.076723

1. To construct cluster groups, we assess the values of the indicators. For this purpose, in the context menu, we need to standardize the initial data. So, select *Fill / Standardize Block / Standardize Columns* as shown in Fig. 6.1.



	1 X1	2 X2	3 X3	4 X4	5 X5
Austria	0,512517	8,139704	28,08643	0,657682	1,229883
Belgium	0,441011	11,18419	46,6101	0,423147	1,198028
Bulgaria	0,390566	67,68311	473,7965	0,728101	0,629
Finland	0,659931	13,74058	0	0,644842	1,181
France	0,598438	7,578963	6,85112	0,746463	0,569
Germany	0,425501	9,358731	2922,118	0,671615	0,998
Italy	0,422215	9,180896	4,465834	0,526936	1,333
Poland	0,447059	27,10571	6933,373	1,057638	1,030
Spain	0,493978	10,29191	4295,369	0,61782	0,923
Sweden	0,828147	8,865035	0	0,909419	0,948
Switzerland	0,694735	4,621731	0	0,75509	0,952
UK	0,389766	7,828034	1005,633	1,954056	2,160
Belorussia	0,603121	80,72323	0	0,521061	1,342
Russia	0,562118	121,3625	3960,709	2,899916	2,942
Ukraine	0,541134	179,2357	2740,338	0,676241	1,078

	1 X1	2 X2	3 X3	4 X4	5 X5
Austria	-0,17076	-0,56815	-0,67541	-0,39902	-0,00755
Belgium	-0,73874	-0,50982	-0,66688	-0,75668	-0,06092
Bulgaria	-1,13942	0,572676	-0,47012	-0,29163	-1,01414
Finland	1,000142	-0,46084	-0,68834	-0,4186	-0,08932
France	0,511704	-0,57889	-0,68519	-0,26363	-1,11444
Germany	-0,86193	-0,5448	0,657549	-0,37777	-0,39445
Italy	-0,88803	-0,5482	-0,68629	-0,5984	0,165506
Poland	-0,6907	-0,20477	2,505087	0,21091	-0,34177
Spain	-0,31802	-0,52692	1,290053	-0,45981	-0,5206
Sweden	2,336278	-0,55425	-0,68834	-0,01512	-0,47917
Switzerland	1,276589	-0,63555	-0,68834	-0,25047	-0,47298
UK	-1,14577	-0,57412	-0,22516	1,577936	1,551189
Belorussia	0,548901	0,822519	-0,68834	-0,60736	0,181907
Russia	0,223215	1,60115	1,135912	3,020361	2,860878
Ukraine	0,05654	2,709977	0,573823	-0,37072	-0,26414

Fig. 6.1. **The normative values of the energy security indices**

2. To perform cluster analysis, we need to log into the cluster analysis module, using the *Statistics / Multivariate Exploratory / Cluster analysis* menu (Fig. 6.2).

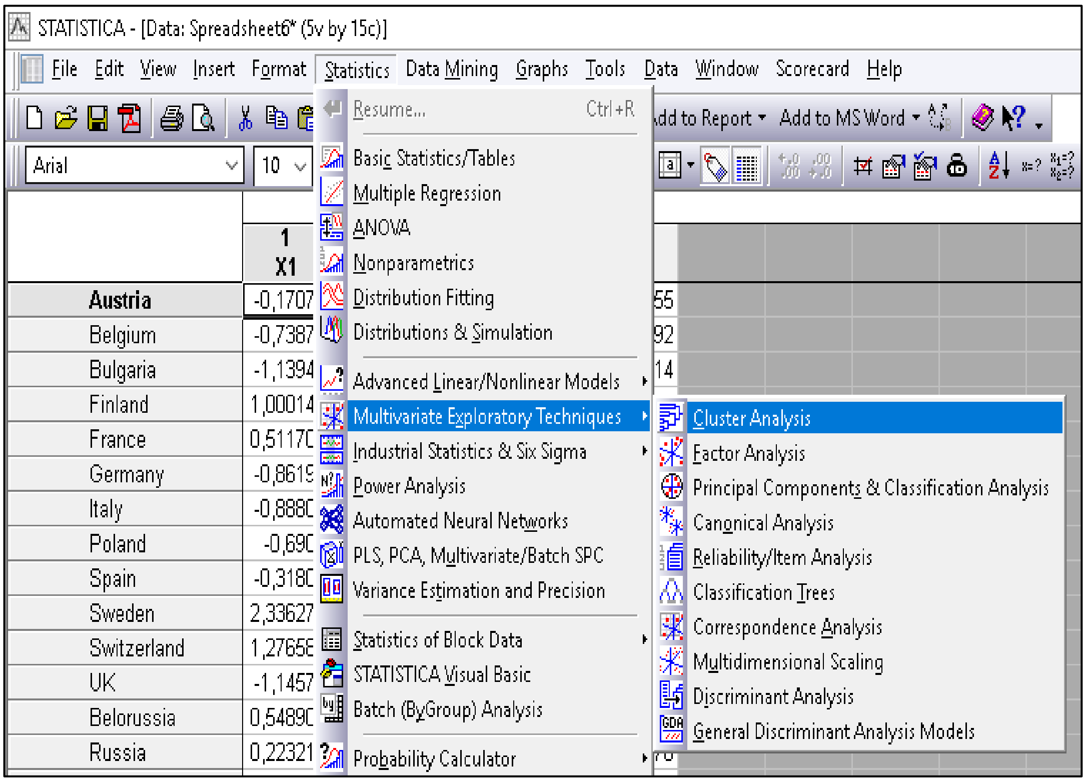


Fig. 6.2. **The cluster analysis module**

In the resulting dialog window you may choose one of the following methods of clustering:

1) *joining (tree clustering)*;

2) *k-means clustering*;

3) *two-way joining*.

Let's start cluster analysis with the method of natural hierarchical clustering – Ward's method (Fig. 6.3).

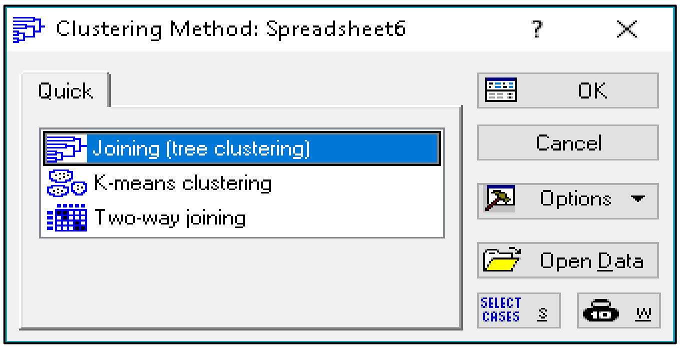


Fig. 6.2. **Choosing the Ward's method**

3. In order to determine the number of clusters, it is expedient to initially conduct natural (tree-like) clustering. In the Statistica package, this type of clustering involves the implementation of several stages.

3.1. Selection of indicators for which clustering is carried out (Fig. 6.4).

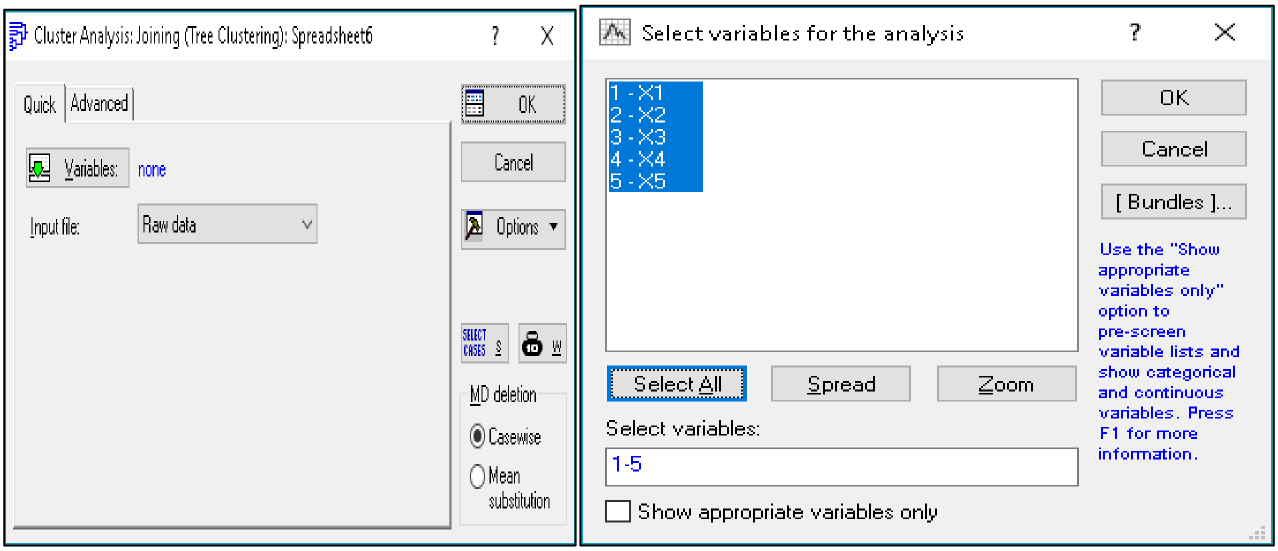


Fig. 6.4. **Selection of the indicators for the analysis**

3.2. Selection of objects of classification in the *Cluster* field. When clustering the variables themselves, they are labeled *Variables* [Columns], in this task – *Cases* [rows] (*Observation* [rows]) (Fig. 6.5).

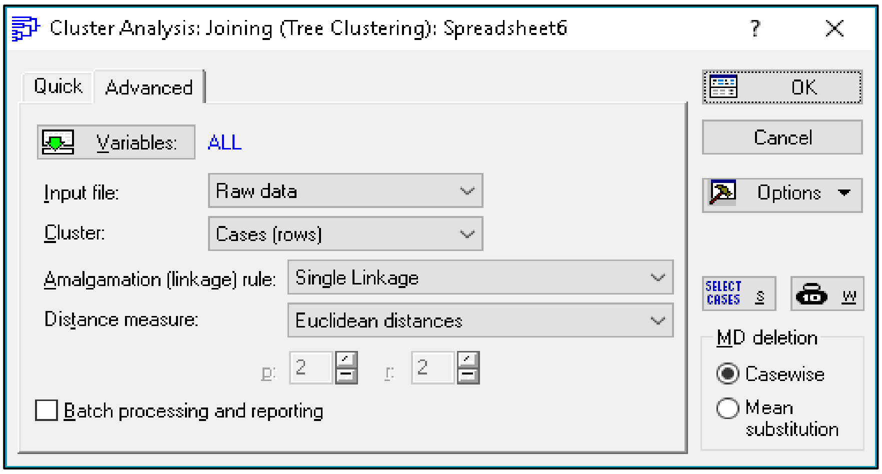


Fig. 6.5. **Choosing the parameters of cluster analysis**

3.3. The choice of rules for grouping the objects. To do this, use the *Amalgamation* [*linkage*] *rule* menu, which allows you to choose one of the following rules:

- *Single linkage* (the one-way method "Closest neighbor's principle");

- ➢ *Complete linkage* (the full-length method);
- ➢ *Unweighted pair-group average* (the unweighted pair average);
- ➢ *Weighted pair-group average* (the weighted pairwise average);
- ➢ *Unweighted pair-group centroid* (the unweighted centroid method);
- ➢ *Weighted pair-group centroid* (the weighted centroid method);
- ➢ *Ward's method*.

According to the work propose, let's use the single-link method.

3.4. Choosing the distance type to be used in the clustering process. For this purpose, in the *Distance measure* window, you must select one of the distance types used in the package:

• *Squared Euclidean distances* (the square of the Euclidean distance);

• *Euclidean distances*;

• *City-block* (*Manhattan*) distance (the distance from the city districts (Manhattan distance));

• *Chebyshev distance metric* (Chebyshev distance);

• *Percent disagreement*.

According to work propose, let's use the Euclidean distance.

After setting all clustering parameters, we go to the window of results (Fig. 6.6).

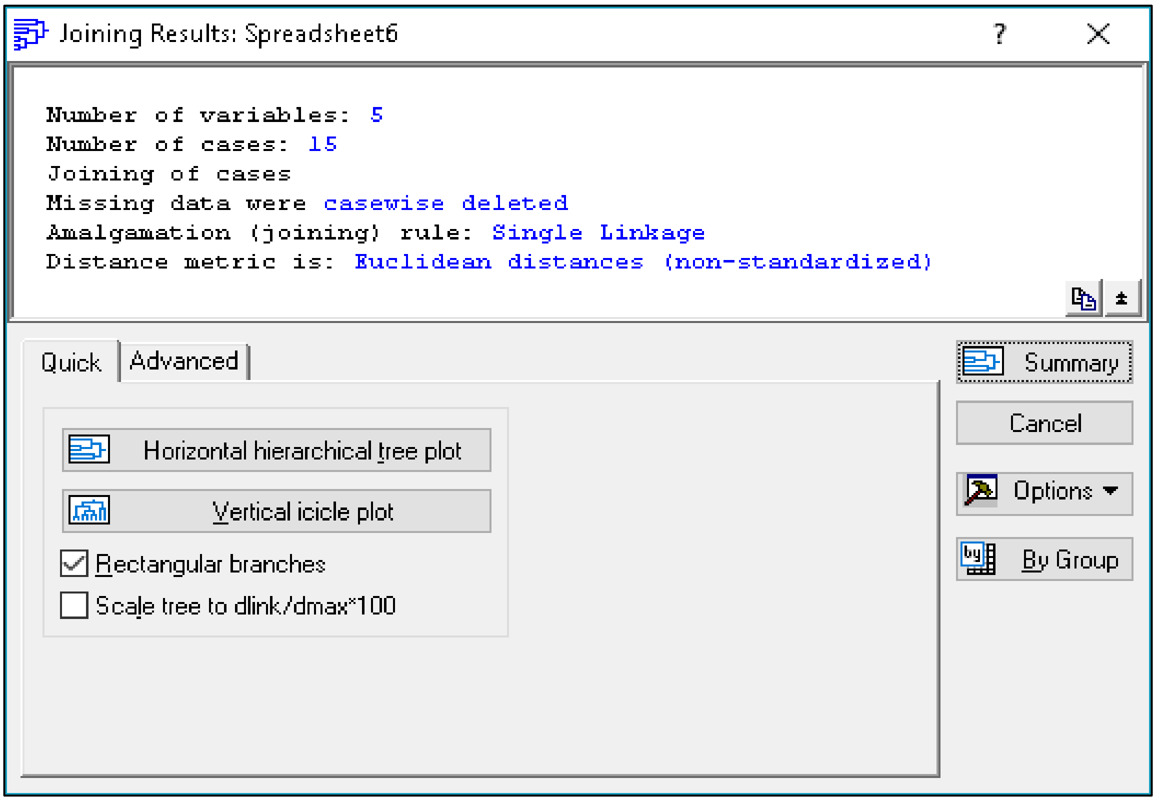


Fig. 6.6. **Choosing the parameters of cluster analysis**

Using the *Vertical (Horizontal) icicle plot* button, we build a vertical dendrogram (Fig. 6.7) and a horizontal hierarchical dendrogram (Fig. 6.8).

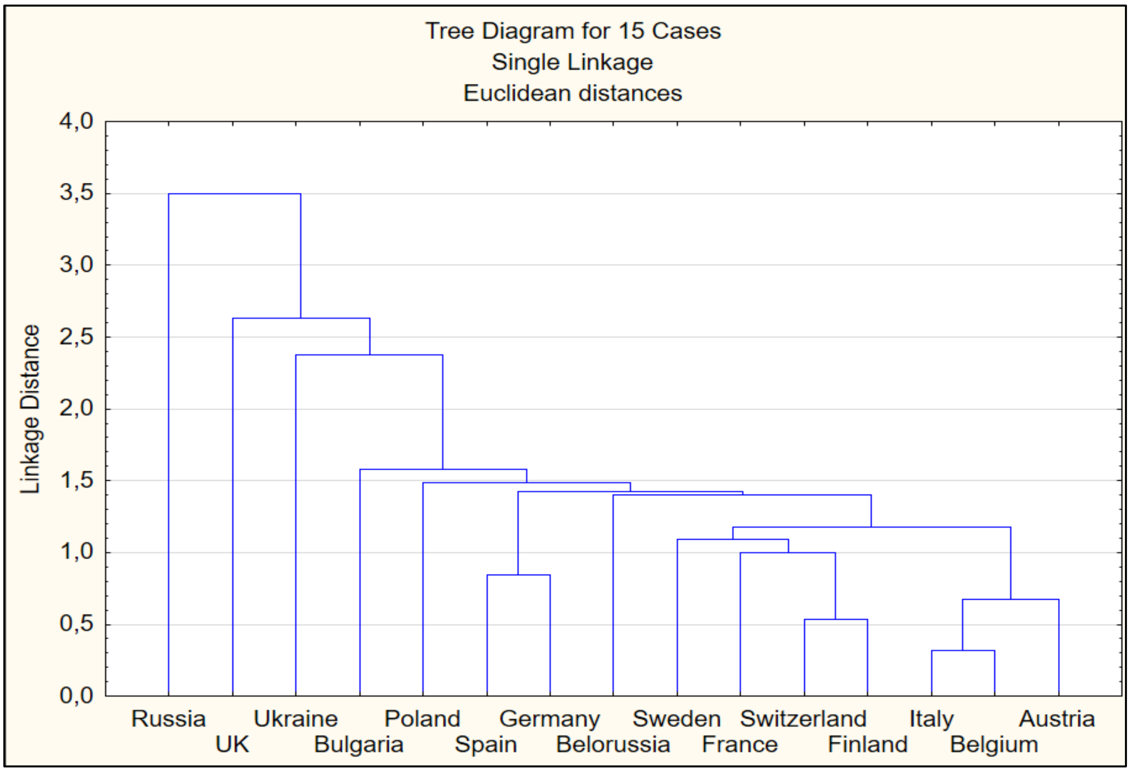


Fig. 6.7. **The vertical dendrogram**

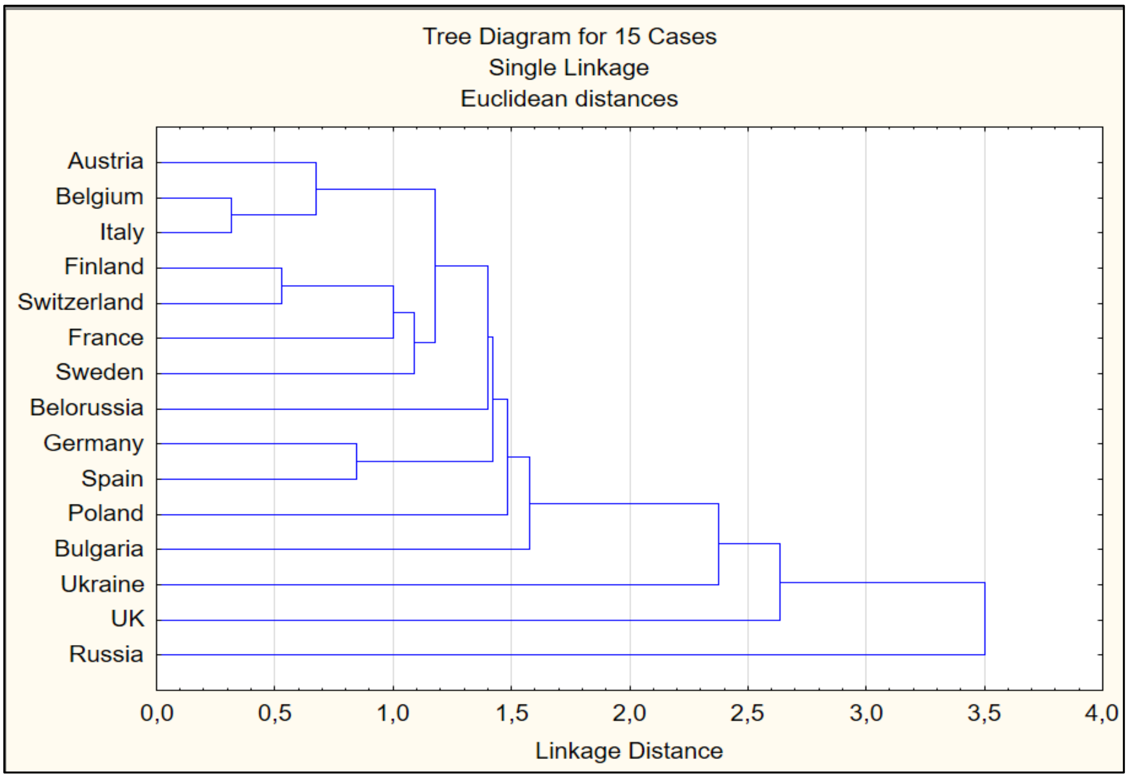


Fig. 6.8. **The horizontal hierarchical dendrogram**

The analysis of the determination of the number of cluster groups is based on the use of dendrograms. The most expedient way is to break the aggregate of countries into 4 clusters.

4. For the confirmation of the results of clusterization by natural methods, clustering is carried out by artificial methods. Construction of clusters using the k-medium method is performed in the following stages:

4.1. Setting up the key clustering parameters. The clustering way and clustering objects are selected similar to the tree-clustering method. Taking into account the results of constructing the dendrogram, the number of clusters is equal to 4 (Fig. 6.9).

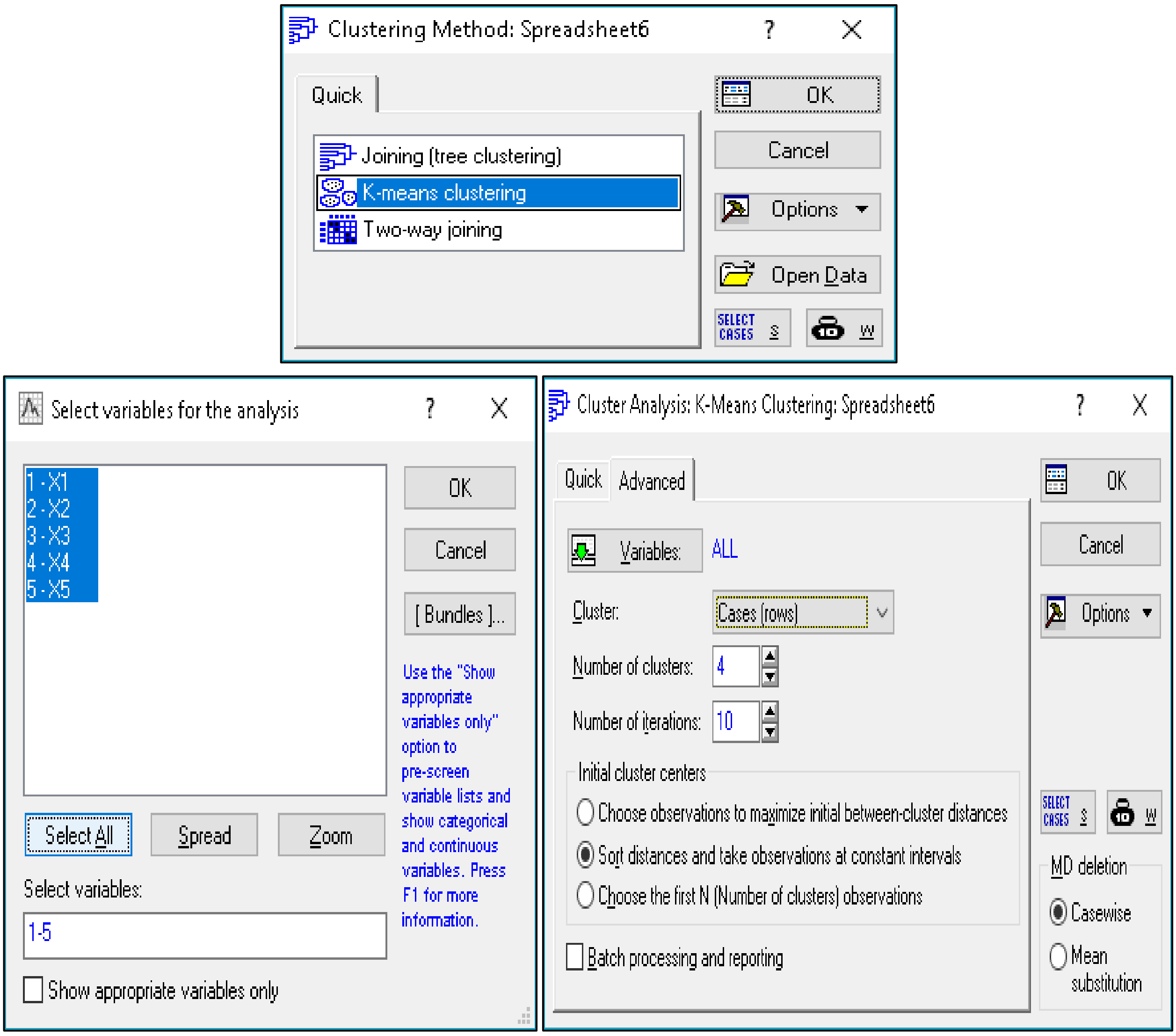


Fig. 6.9. **The stages of the k-means method**

4.2. In the *Clustering results* window, you can select those calculations and reports for the cluster analysis that the user needs (Fig. 6.10).

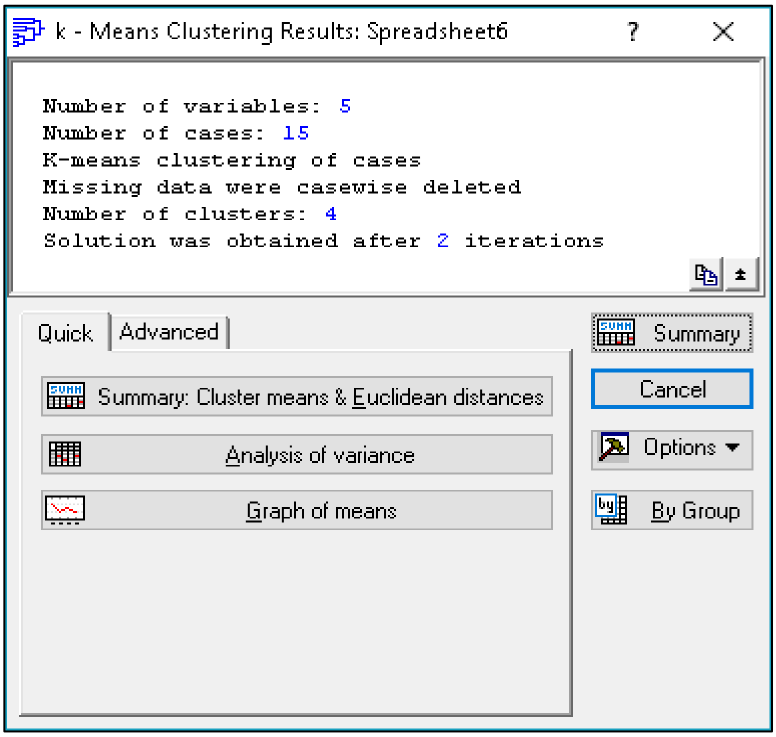


Fig. 6.10. **Selection of the parameters for the k-means method**

4.3. Let's consider the stages of the k-means method.

4.3.1. Use the *Cluster Means & Euclidean Distances* button (mean values in clusters and Euclidean distances) (Fig. 6.11).

Cluster Number	Euclidean Distances between Clusters (Spreadsheet6) Distances below diagonal Squared distances above diagonal **No. 1**	No. 2	No. 3	No. 4
No. 1	0,000000	3,060827	0,748075	0,914698
No. 2	1,749522	0,000000	3,612635	2,797337
No. 3	0,864913	1,900693	0,000000	1,520397
No. 4	0,956398	1,672524	1,233044	0,000000

Fig. 6.11. **Euclidean distances**

Based on the matrix of distances between the clusters, one can determine the quality of the clusterization carried out. The greater the distance between clusters and the less the distance between the elements of the clusters, the more qualitative clustering is carried out.

4.3.2. The *Descriptive Statistics* button for each cluster allows you to define descriptive statistics for each cluster (Fig. 6.12).

Variable	Descriptive Statistics for Cluster 1 (Spreadsheet6) Cluster contains 5 cases		
	Mean	Standard Deviation	Variance
X1	-0,759776	0,359997	0,129598
X2	-0,319659	0,499272	0,249272
X3	-0,368229	0,580369	0,336828
X4	-0,484700	0,189046	0,035738
X5	-0,262310	0,466688	0,217798

Variable	Descriptive Statistics for Cluster 2 (Spreadsheet6) Cluster contains 2 cases		
	Mean	Standard Deviation	Variance
X1	-0,461279	0,968021	0,937065
X2	0,513514	1,538150	2,365906
X3	0,455375	0,962425	0,926261
X4	2,299148	1,019948	1,040294
X5	2,206034	0,926090	0,857643

Variable	Descriptive Statistics for Cluster 3 (Spreadsheet6) Cluster contains 5 cases		
	Mean	Standard Deviation	Variance
X1	1,134723	0,743860	0,553328
X2	-0,281405	0,620325	0,384803
X3	-0,687714	0,001411	0,000002
X4	-0,311037	0,219579	0,048215
X5	-0,394801	0,489137	0,239255

Variable	Descriptive Statistics for Cluster 4 (Spreadsheet6) Cluster contains 3 cases		
	Mean	Standard Deviation	Variance
X1	-0,317392	0,373618	0,139591
X2	0,659430	1,783115	3,179501
X3	1,456321	0,976309	0,953179
X4	-0,206537	0,364254	0,132681
X5	-0,375504	0,131519	0,017297

Fig. 6.12. **The descriptive statistics for each cluster**

4.3.3. The list of countries included in each cluster can be obtained using the *Members for each cluster & distances* button (group members and distances) (Fig. 6.13).

	Members of Cluster Number 1 and Distances from Respectiv Cluster contains 5 cases
	Distance
Austria	0,339201
Belgium	0,219238
Bulgaria	0,557367
Germany	0,477962
Italy	0,270480

	Members of Cluster Number 2 and Distances from Respectiv Cluster contains 2 cases
	Distance
UK	0,782762
Russia	0,782762

	Members of Cluster Number 3 and Distances from Respective Cluster contains 5 cases
	Distance
Finland	0,176180
France	0,446494
Sweden	0,567955
Switzerland	0,176254
Belorussia	0,629641

	Members of Cluster Number 4 and Distances from Respective Cluster contains 3 cases
	Distance
Poland	0,657497
Spain	0,551409
Ukraine	1,016142

Fig. 6.13. **The members of each cluster**

Fig. 6.13 shows that the representative for the first cluster is Belgium, for the second – Russia and the United Kingdom, for the third cluster – Finland, for the fourth one – Spain. A comparative analysis of the Euclidean distances allowed us to conclude that the built-up clusterization is qualitative, as evidenced by a significant excess of distance between the groups and within them.

4.3.4. To plot a graph showing the pattern of the breakdown of countries into clusters depending on the level of energy security, the *Graph of means* button is used (Fig. 6.14).

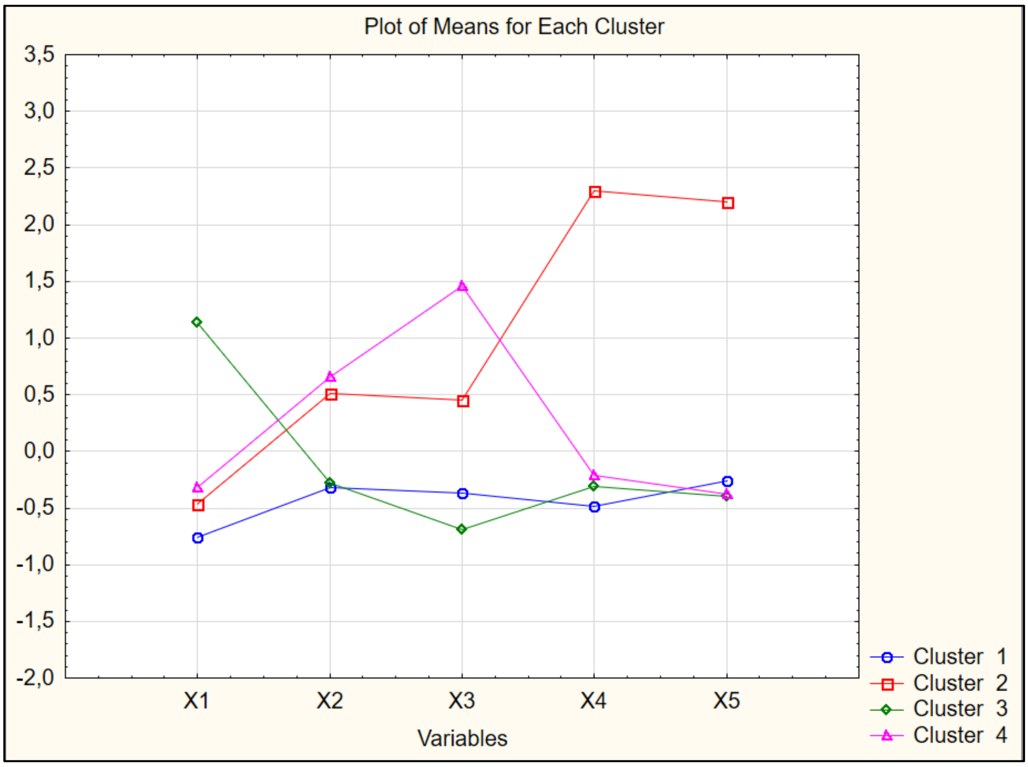


Fig. 6.14. **The graphs of average values of indicators for 4 clusters**

Analyzing the results we can draw conclusions and characterize the above clusters (Table 6.2).

Table 6.2

The general characteristics of the energy security clusters

Cluster number	List of countries included in the cluster	Key characteristics of the class	Recommendation
1	2	3	4
The first cluster	Austria, Belgium, Germany, Bulgaria, Italy	Countries with an average level of energy resources and high energy-saving technologies	Diversify suppliers of energy resources

Table 6.2 (the end)

1	2	3	4
The second cluster	Russia, the UK	Absolutely energy independent countries with large own energy reserves but inefficient use of energy resources	Increase energy efficiency
The third cluster	Finland, France, Sweden, Switzerland, Belarus	Countries with low energy resources and high levels of energy saving technologies	Diversify energy suppliers, use non-traditional sources of energy resources
The fourth cluster	Poland, Spain, Ukraine	Countries with an average level of energy resources and very low energy-saving technologies	Diversify energy suppliers, increase energy efficiency

Thus, we can conclude that Ukraine is among the countries with an average level of energy supply and very low energy-saving technologies.

Topic 7. Data recognition and discriminant analysis

Laboratory work 7
Solving the problem of classification by the method of discriminant analysis

The purpose of the work is to obtain skills in the use of discriminant analysis in the Statistica package.

The task is to check the quality of clustering of the countries of the world in terms of energy security.

Guidelines

The initial data about 15 countries of the world, which were distributed in four groups by the method of cluster analysis (in terms of energy security), are shown in Fig. 7.1.

	1 X1	2 X2	3 X3	4 X4	5 X5	6 # Cluster
Austria	-0,17076	-0,56815	-0,67541	-0,39902	-0,00755	1
Belgium	-0,73874	-0,50982	-0,66688	-0,75668	-0,06092	1
Bulgaria	-1,13942	0,572676	-0,47012	-0,29163	-1,01414	1
Finland	1,000142	-0,46084	-0,68834	-0,4186	-0,08932	3
France	0,511704	-0,57889	-0,68519	-0,26363	-1,11444	3
Germany	-0,86193	-0,5448	0,657549	-0,37777	-0,39445	1
Italy	-0,88803	-0,5482	-0,68629	-0,5984	0,165506	1
Poland	-0,6907	-0,20477	2,505087	0,21091	-0,34177	4
Spain	-0,31802	-0,52692	1,290053	-0,45981	-0,5206	4
Sweden	2,336278	-0,55425	-0,68834	-0,01512	-0,47917	3
Switzerland	1,276589	-0,63555	-0,68834	-0,25047	-0,47298	3
UK	-1,14577	-0,57412	-0,22516	1,577936	1,551189	2
Belorussia	0,548901	0,822519	-0,68834	-0,60736	0,181907	3
Russia	0,223215	1,60115	1,135912	3,020361	2,860878	2
Ukraine	0,05654	2,709977	0,573823	-0,37072	-0,26414	4

Fig. 7.1. **The initial data with cluster distribution**

Discriminant analysis is a multidimensional statistical method that allows one to study the differences between two or more groups of objects in several variables at a time. The main task of discriminant analysis is to study group differences, that is, to discriminate objects based on certain attributes.

The choice of this module is possible through the *Statistics / Multivariate Exploratory Techniques / Discriminant analysis* menu (Fig. 7.2).

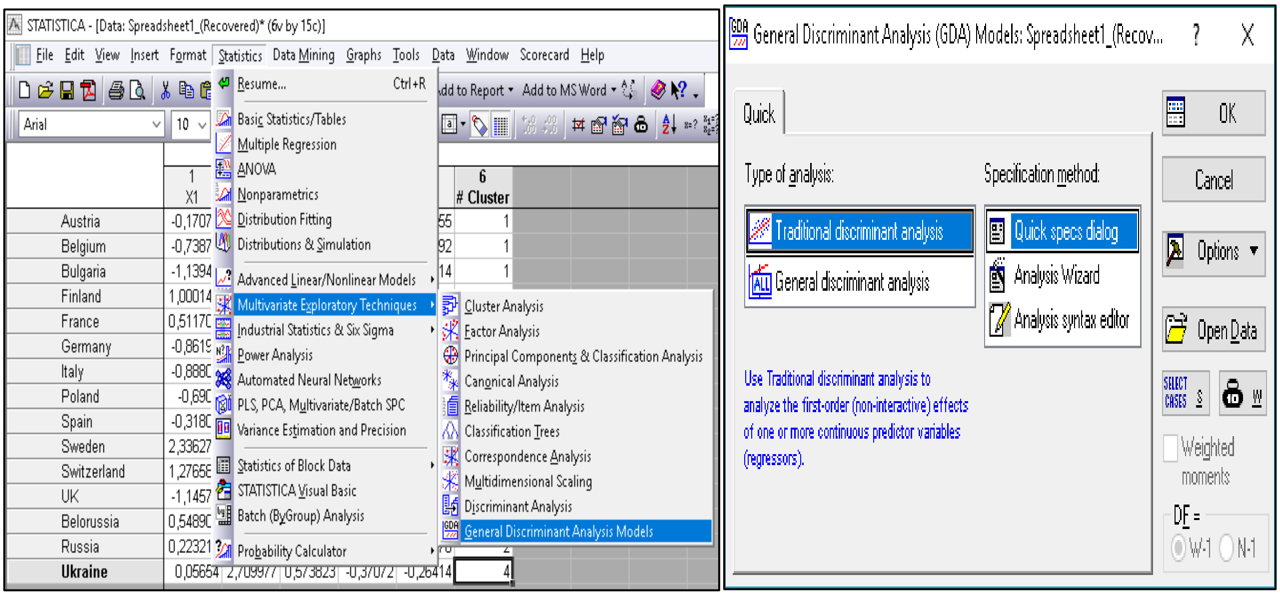


Fig. 7.2. **The first way to choose discriminant analysis**

Or, perhaps, use the *Module Switcher* tab, which contains a list of all available modules, and press *Discriminant Analysis* and then the *Switch to* button (Fig. 7.3).

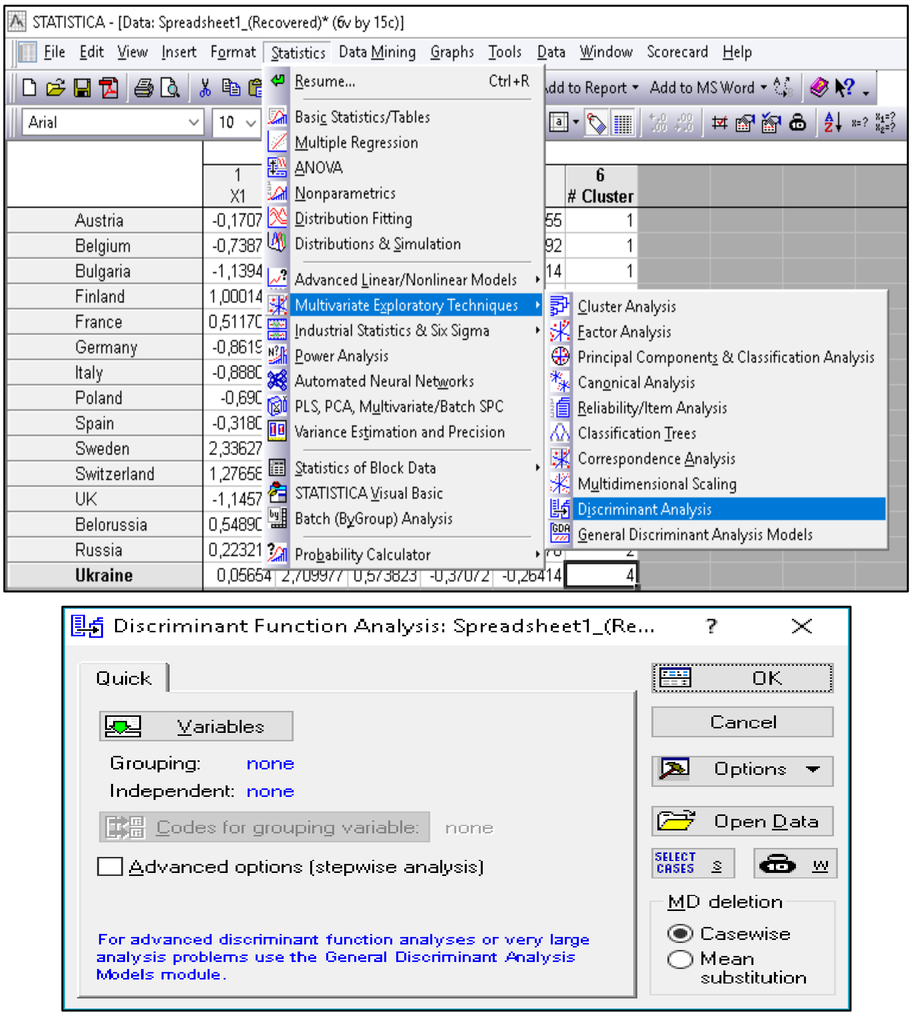


Fig. 7.3. **The second way to choose discriminant analysis**

The screen for the *Discriminant Function Analysis* module will appear, with the help of which you can perform the following functions:

- open the data file using the *Open Data* button;
- select the variable – *Variables*;
- determine the number of groups of objects being analyzed – *Codes for grouping variables*;
- permanently remove variables from the *Casewise* list or replace them with average *Mean substitution*;
- specify the conditions for selecting observations from the database – *Select Cases*;
- add the share (weight) of variables by selecting them from the list – *W*.

You can use the *Variables* button to select a *Grouping* and an *Independent variable* (Fig. 7.4).

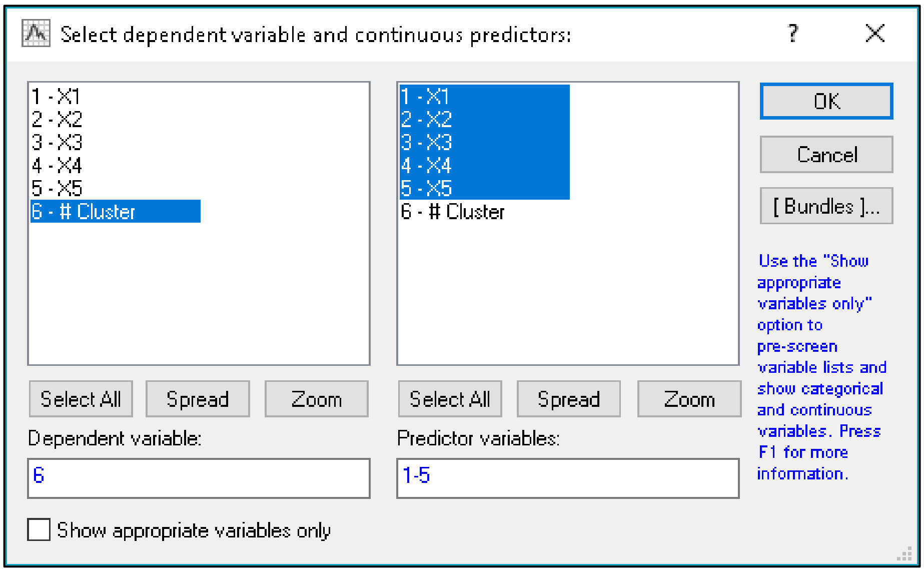


Fig. 7.4. **Selection of variables**

Using the buttons located in the variables selection panel you can:

- select all variables – *Select All*;
- view the type of name – *Spread*;
- see additional information on the *Zoom* variable.
- define the model by clicking the OK button.

The *Model Definition* dialog box that is used to select a model is shown in Fig. 7.5.

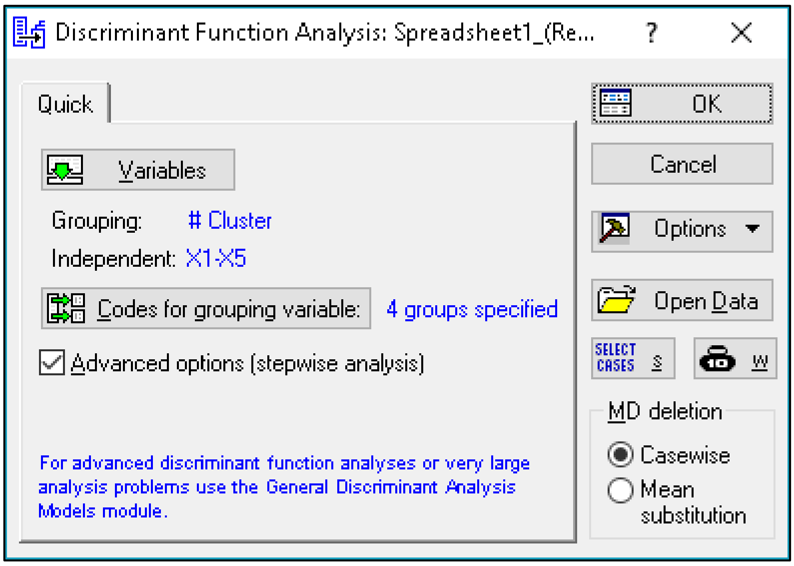


Fig. 7.5. **Choosing the parameters of the discriminant analysis**

On the *Advanced* tab, you can specify the method that will be used to select meaningful variables (Fig. 7.6).

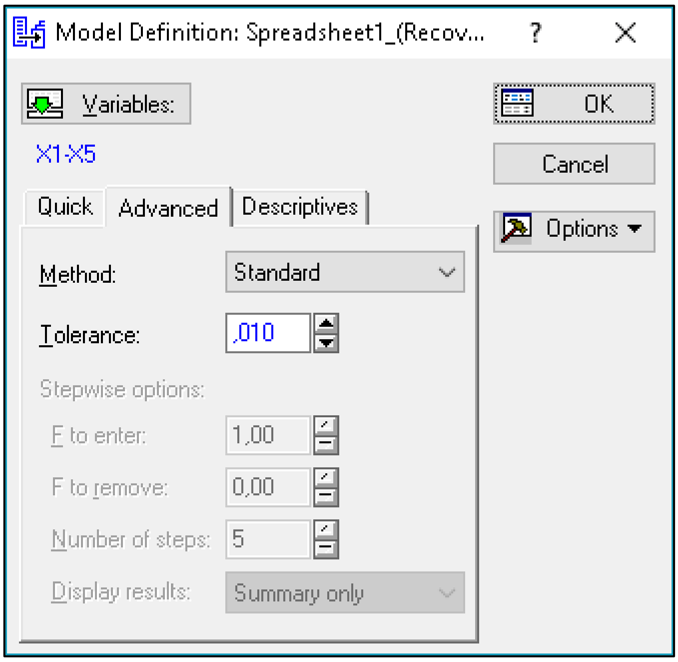


Fig. 7.6. **Choosing the method of discriminant analysis**

The following methods may be used:

• *Standard.* All variables are included in the model at the same time;

• *Forward stepwise.* At each step, in the model, a variable with a maximum F value is selected. The procedure ends when all variables whose values F are greater than the values specified in the F field to enter are included in the model;

• *Backward stepwise* (step-by-step). At each step, all variables are selected in the model, which are then deleted depending on the value of F. The steps end when there are no variables with F values less than certain specified by the user in the F to remove the field.

The *Number of steps* field determines the maximum number of analysis steps that the procedure ends in.

The *Tolerance* field allows you to exclude non-informative variables from the model. If the tolerance is less than the value of 0.01, the variable is considered non-informative and should not be included in the model.

Unlike the standard method for step-by-step procedures, there are two modes (output of results) of the analysis of *Display of results*:

• *At every step* – the program displays the results dialog, received at each step, starting from zero.

• *Summary only* (in the final step) displays a result window only in the last step, but it contains the option to view the main summary statistics for the step-by-step procedure as well.

Descriptive / Review Descriptive Statistics allows you to get descriptive statistics for selected variables:

✓ *Pooled within-group covariances & correlations* (combined intragroup covariations and correlations);

✓ *Total covariances & correlations* (full covariance and correlations);

✓ *Graph* (graphs of correlation functions for all variables);

✓ *Means & number of cases* (mean values for each variable);

✓ *Box & scale* diagrams;

✓ *Standard deviations* (standard deviations of variables in each group);

✓ *Categorized histogram* (by group) (histograms categorized in groups for each variable);

✓ *Box & whisker* plot (by group) (scale diagrams by groups);

✓ *Categorized scatterplot* (by group) for two of any variables;

✓ *Categorized normal probability* plot (normal graphs categorized for any variable in groups).

As a method of analysis, choose *Standard* based on the results obtained in the computations presented in the *Discriminant Function Analysis Results* window (Fig. 7.7).

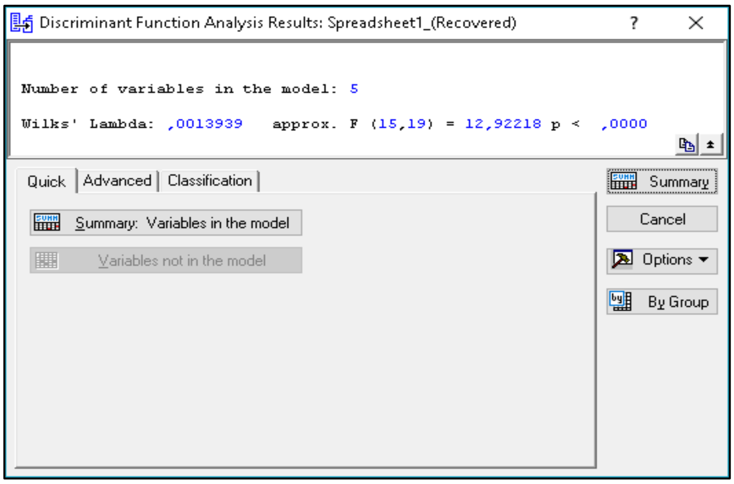


Fig. 7.7**. The *Discriminant function analysis results* window**

It is possible to obtain the following information:

• the number of variables in the model – 5;

• the value of Wilks' lambda which is 0.0014;

- the approximate value of F-statistics related to Wilks' lambda (Approx. F (15, 19)) – 12,922.

The level of significance of F is the criterion $p < 0.0000$ for the value 12,922.

The values of Wilks' statistics are in the range 0 – 1. If Wilks' statistics are close to 0, this indicates good discrimination, while the values close to 1 indicate bad discrimination.

Consequently, according to Wilks' lambda, which is 0.0014 and F criterion equal to 12.922 ($F_{table} < F_{calc}$), it is possible to conclude that the classification is correct.

As a validation check, see the results of the classification matrix by clicking the *Classification matrix* button (Fig. 7.8), pre-selecting *Same for all groups* in the right-hand window of *Discriminant Function Analysis* Results (Fig. 7.9).

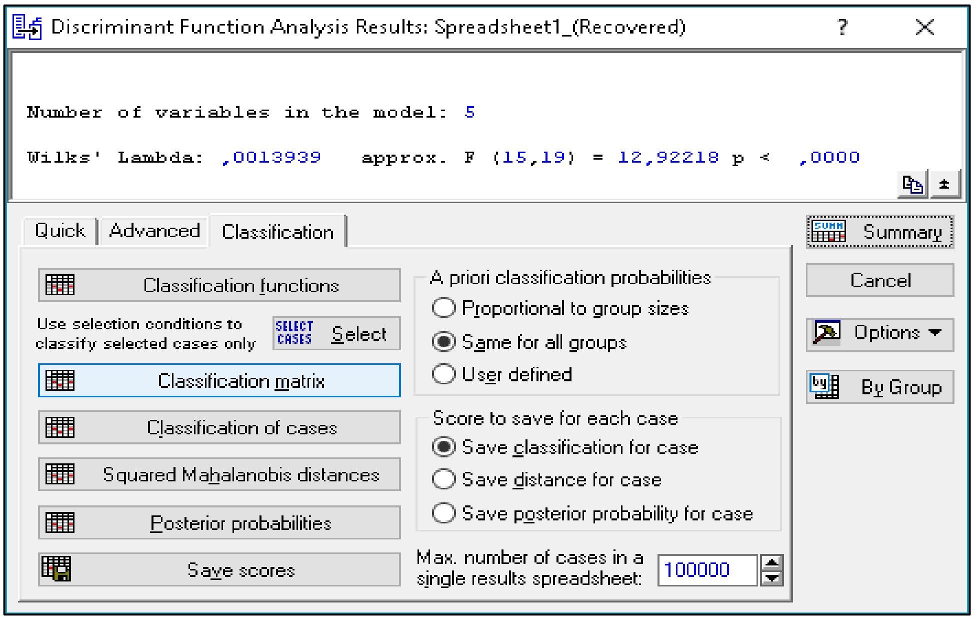


Fig. 7.8. ***The Classification matrix* button**

Classification Matrix (Spreadsheet1_(Recovered))
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
G_1:1	100,0000	5	0	0	0
G_2:2	100,0000	0	2	0	0
G_3:3	100,0000	0	0	5	0
G_4:4	100,0000	0	0	0	3
Total	100,0000	5	2	5	3

Fig. 7.9. **The Classification matrix**

Based on the results of the classification matrix, we can conclude that the objects are correctly broken down into four groups by cluster analysis. If there are countries that are incorrectly assigned to the appropriate groups, see *Classification of cases* (Fig. 7.10).

Case	Classification of Cases (Spreadsheet1_(Recovered)) Incorrect classifications are marked with * Observed Classif.	1 p=,25000	2 p=,25000	3 p=,25000	4 p=,25000
Austria	G_1:1	G_1:1	G_3:3	G_4:4	G_2:2
Belgium	G_1:1	G_1:1	G_3:3	G_4:4	G_2:2
Bulgaria	G_1:1	G_1:1	G_3:3	G_4:4	G_2:2
Finland	G_3:3	G_3:3	G_1:1	G_4:4	G_2:2
France	G_3:3	G_3:3	G_1:1	G_4:4	G_2:2
Germany	G_1:1	G_1:1	G_4:4	G_3:3	G_2:2
Italy	G_1:1	G_1:1	G_3:3	G_4:4	G_2:2
Poland	G_4:4	G_4:4	G_1:1	G_3:3	G_2:2
Spain	G_4:4	G_4:4	G_3:3	G_1:1	G_2:2
Sweden	G_3:3	G_3:3	G_1:1	G_4:4	G_2:2
Switzerland	G_3:3	G_3:3	G_1:1	G_4:4	G_2:2
UK	G_2:2	G_2:2	G_1:1	G_3:3	G_4:4
Belorussia	G_3:3	G_3:3	G_1:1	G_4:4	G_2:2
Russia	G_2:2	G_2:2	G_1:1	G_3:3	G_4:4
Ukraine	G_4:4	G_4:4	G_3:3	G_1:1	G_2:2

Fig. 7.10. **The classification of cases**

In this figure incorrectly assigned objects are marked with an asterisk (*). Thus, the task of getting the correct groups is complete.

The classification of the cases of training samples aims to exclude from the training samples those objects that by their indicators do not correspond to most of the objects forming a homogeneous group. To do this, the metric of Mahalanobis defines the distance from all n objects to the center of gravity of each group (the vector of averages), which are determined by the training sample. The assignment of the *i*-th object to the *j*-th group is considered false if the distance of Mahalanobis from the object to the center of its group is much longer than from the object to the center of other groups, and the a posteriori probability of falling into its group is below the critical value. In this case, the object is considered incorrectly assigned and should be excluded from the sample.

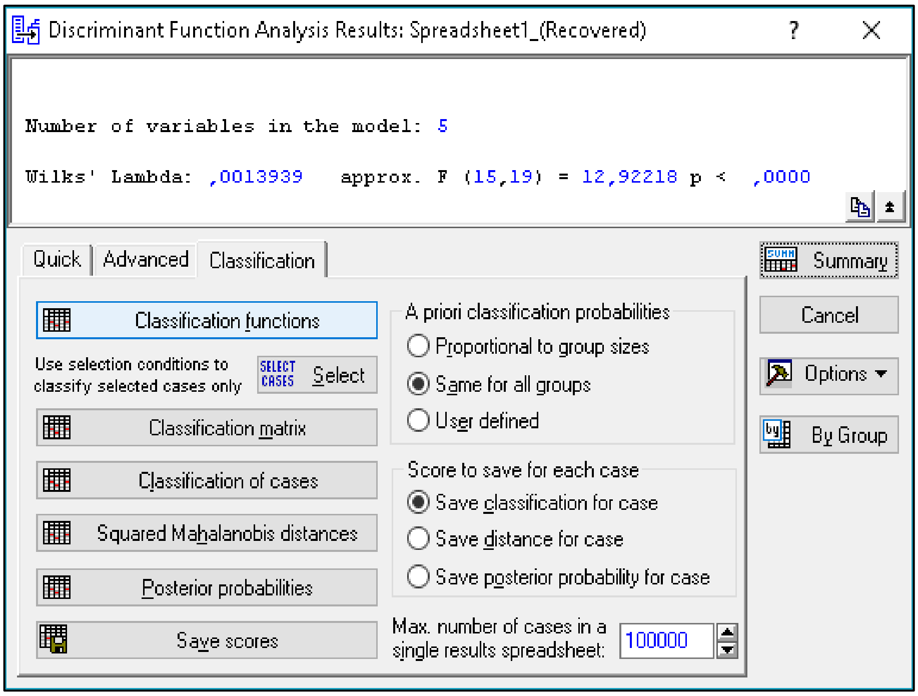
The exclusion of the object from the training samples means that in the table of output data in the object which should be excluded from the sample (it is marked with "*"), the number of belonging to this group is removed, after which the testing process is repeated. By assumption, at first the object which is least suitable to a certain group is chosen, that is, the object for which the greatest distance is Mahalanobis and a posteriori probability is the least.

When removing another object from a group it is necessary to remember that at the same time the center of gravity of the group (the vector of averages) is shifted since it is determined by the rest of the observations. After removing the next item from the list of training samples, it is possible that new incorrectly assigned objects will appear, which, prior to deletion, have been taken into account as being properly classified. Therefore, this procedure should be carried out by deleting only one object at each step and returning it to the training sample.

The procedure for excluding observations continues as long as the overall correlation coefficient in the classification matrix reaches 100 %, that is, all observations of the training samples will be correctly assigned to the corresponding groups. The results of the received training samples are presented in the *Discriminant Function Analisis Results* window. As a result of the analysis, the total coefficient of correctness of the training samples should be 100 %.

On the basis of the received training samples, it is possible to conduct a reclassification of those objects that are not included in the training sample and any other objects subject to the grouping, therefore it is necessary to specify the basis of classification functions.

To do this, in the *Discriminant Function Analysis Results* window, click *Classification functions* (Fig. 7.11). A window will appear, from which it is possible to write the classification functions for each class.



	Classification Functions; grouping: # Cluster (Spreadsheet1_(Recovered))			
Variable	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
X1	-1,47918	-33,365	8,6568	10,2805
X2	0,02172	-11,154	1,1139	5,5432
X3	-0,10984	-41,812	4,7592	20,1256
X4	-1,99439	86,011	-13,1602	-32,0826
X5	-0,01412	25,143	-5,0471	-8,3268
Constant	-2,45016	-123,307	-7,5476	-21,1137

Fig. 7.11. **Classification functions**

Countries with an average level of energy resources and high energy-saving technologies = - 1.479X1 + 0.0217X2 - 0.1098X3 - 1.994X4 - 0.014X5.

Absolutely energy independent countries with large own energy reserves but inefficient use of energy resources = - 33.365X1 - 11.154X2 - 41.812X3 + + 86.011X4 + 25.143X5.

Countries with low energy resources and high levels of energy saving technologies = 8.657X1 + 1.1139X2 + 4.759X3 - 13.16X4 - 5.047X5.

Countries with an average level of energy resources and very low energy-saving technologies = 10.281X1 + 5.543X2 + 20.123X3 - 32.08X4 - - 8.327X5.

With these functions, it will be possible to further classify new cases. New cases will refer to the class for which the classification value will be maximal. The choice of the method of final classification depends on the number of new objects subject to classification. If the number of new cases is small, you can apply a method based on statistical criteria. If the number of new cases is large, it is more rational than the training samples to obtain classification functions and then, define the formulas and hold the final classification.

For more detailed information, it is possible to review the results of a canonical analysis that can be performed if at least three groups have been selected and at least two variables in the model are selected by clicking the *Perform canonical analysis* button (Fig. 7.12).

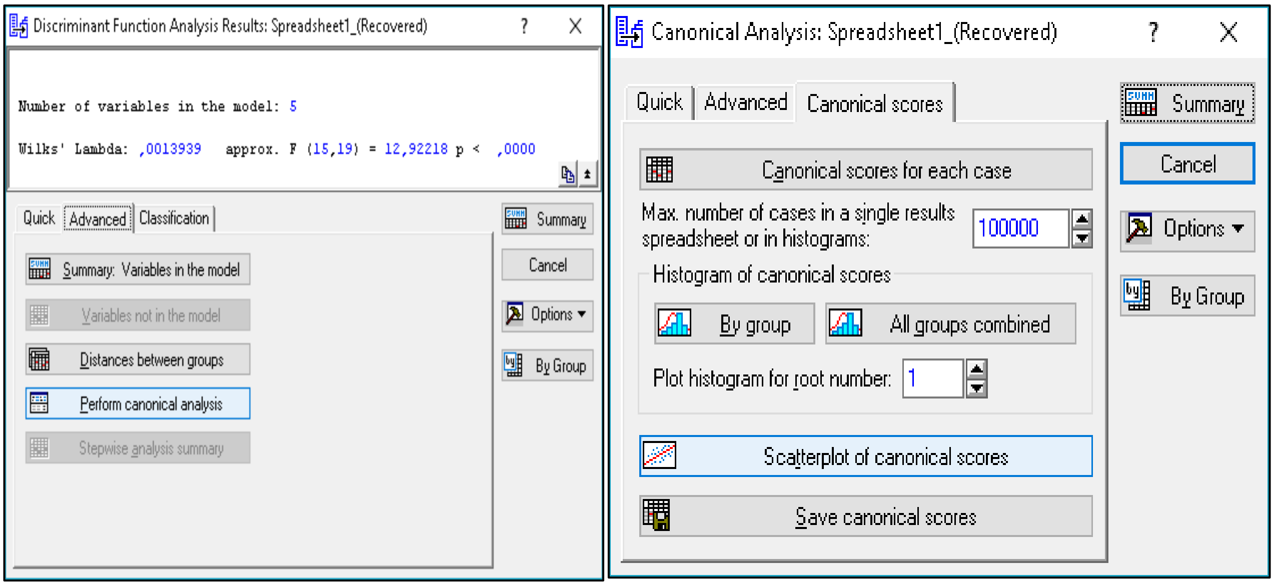


Fig. 7.12. **Choosing the canonical analysis**

A *canonical analysis* window appears in which the *Scatterplot of canonical scores* option is used to construct the next scatterplot for values. With this diagram, it is possible to determine the contribution that each discriminating function places in the distribution between the groups (Fig. 7.13).

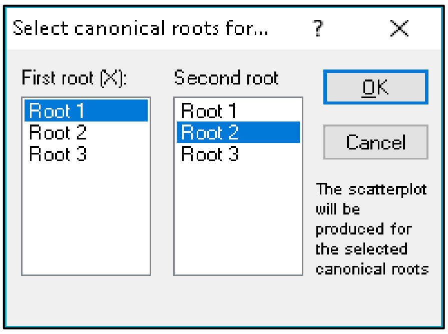


Fig. 7.13. **Selection of parameters of the canonical analysis**

The graph of scattering of canonical values for canonical roots is presented in Fig. 7.14.

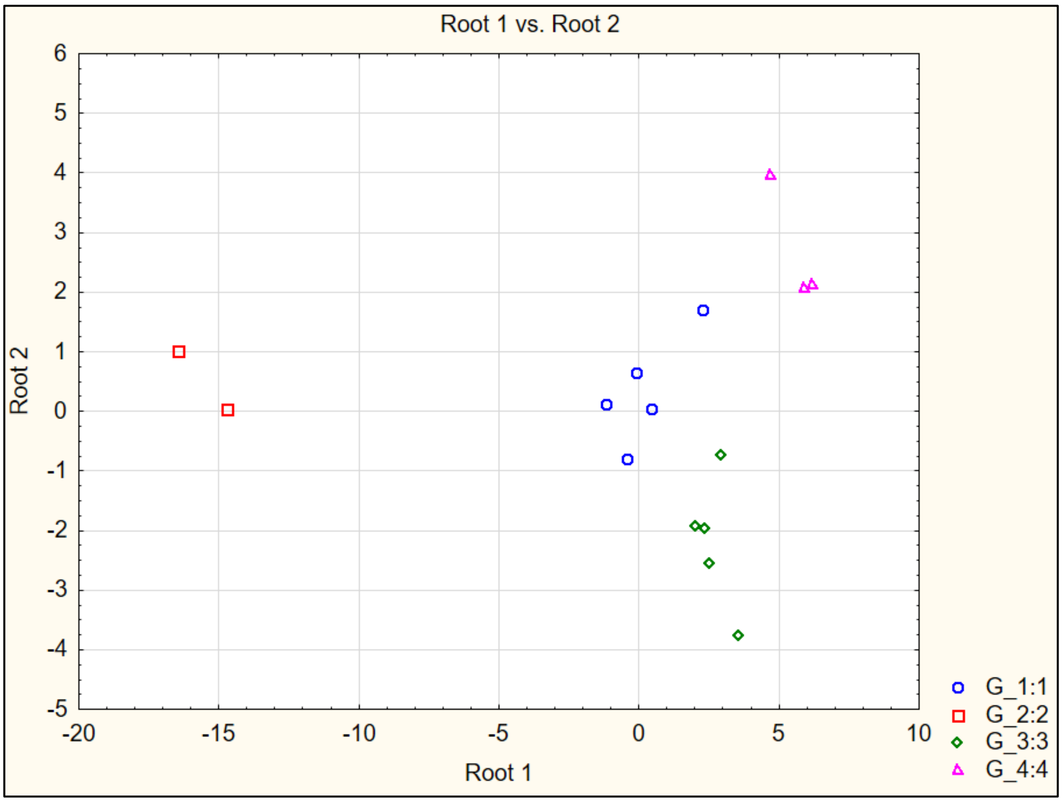


Fig. 7.14. **Visualization of the canonical analysis**

Conclusion. Thus, the classification of countries depending on the level of energy security by the cluster analysis method is adequate. In the course of discriminant analysis, the functions that can be used in the future for assigning a particular country to one of the 4 classes are constructed.

Bibliograhy

Main

1. Антохонова И. В. Методы прогнозирования социально-экономических процессов : учебное пособие / И. В. Антохова. – Улан-Удэ : Изд-во ВСГТУ, 2004. – 212 с.

2. Бабешко Л. О. Основы эконометрического моделирования : учебное пособие / Л. О. Бабешко. – Изд. 3-е. – Москва : Комкнига, 2007. – 432 с.

3. Вітлинський В. В. Моделювання економіки : навч. посіб. / В. В. Вітлинський. – Київ : КНЕУ, 2003. – 408 с.

4. Геєць В. М. Моделі і методи соціально-економічного прогнозування : підручник / В. М. Геєць, Т. С. Клебанова, О. І. Черняк, А. В. Ставицький та інші. – 2-ге вид., виправ. – Харків : ВД "ІНЖЕК", 2008. – 396 с.

5. Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. / А. М. Єріна. – Київ : КНЕУ, 2001. – 170 с.

6. Клебанова Т. С. Эконометрия : учебно-методическое пособие для самостоятельного изучения дисциплины / Т. С. Клебанова, Н. А. Дубовина, Е. В. Раевнева. – Харьков : ИД "ИНЖЭК", 2003. – 132 с.

7. Магнус Я. Р. Эконометрика. Начальный курс : учеб. / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – 8-е изд., испр. – Москва : Дело, 2007. – 504 с.

8. Присенко Г. В. Прогнозування соціально-економічних процесів : навч. посіб. / Г. В. Присенко, Є. І. Равікович. – Київ : КНЕУ, 2005. – 378 с.

9. Статистика : навчальний посібник / під ред. д-ра екон. наук, професора Раєвнєвої О. В. – Харків : Вид-во ХНЕУ, 2010. – 520 с.

10. Сошникова Л. А. Многомерный статистический анализ в экономике : учеб. пособ. для вузов / Л. А. Сошникова, В. Н. Тамашевич, Г. Уебе, М. Шефер ; под ред. проф. В. Н. Тамашевича. – Москва : ЮНИТИ-ДАНА, 1999. – 598 с.

Additional

11. Андрієнко В. Ю. Статистичні індекси в економічних дослідженнях / В. Ю. Андрієнко. – Київ : 2004. – 536 с.

12. Дуброва Т. А. Факторный анализ с использованием пакета STATISTICA : учебное пособие / Т. А. Дуброва, Д. Э. Павлов, Н. П. Осипова ; МГУ экономики, статистики и информатики. – Москва : Финансы и статистика. – 2002. – 598 с.

13. Клебанова Т. С. Моделирование экономики : учебное пособие / Т. С. Клебанова, В. А Забродский, О. Ю. Полякова, В. Л. Петренко. – Харьков : Изд-во ХГЭУ, 2001. – 140 с.

14. Орлов А. И. Организационно-экономическое моделирование : учебник : в 3 ч. / А. И. Орлов. – Москва : Изд-во МГТУ им. Н. Э. Баумана. – 2009. – 254 с.

15. Халафян А. А. STATISTICA 6. Статистический анализ данных / А. А. Халафян. – Москва : ООО "Бином-Пресс", 2008. – 512 с.

16. Dickey D. A. Distribution of the estimators for autoregressive time-series with a unit root / D. A. Dickey, W. A. Fuller // Journal of the American Statistical Association. – 1979. – Vol. 74. – P. 427–431.

17. Granger C. W. J. Forecasting economic time series / C. W. J. Granger, P. Newbold. – 2nd ed. – New York : Academic Press, 1986. – 324 p.

18. Lachenbruch P. A. Discriminant Analysis / P. A. Lachenbruch. – New York : Hafner, 1974 – 234 p.

Internet resources

19. Офіційний сайт державної служби статистики України [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua.

20. Электронный учебник StatSoft [Електронний ресурс]. – Режим доступа : http: // www.statsoft.ru.

Content

Introduction ... 3
Content module 1. The methodological bases of statistical modeling and forecasting ... 4
Topic 1. The categorical basis of statistical modeling and forecasting ... 4
Laboratory work 1 ... 4
Topic 2. Regression models as a means of researching economic processes ... 17
Laboratory work 2 ... 17
Laboratory work 3 ... 33
Topic 3. Modeling and forecasting the development of trends ... 43
Laboratory work 4 ... 43
Content module 2. Modeling and forecasting of multidimensional processes ... 61
Topic 5. Factor analysis of data ... 61
Laboratory work 5 ... 61
Topic 6. Cluster analysis as a means of forming homogeneous data groups ... 71
Laboratory work 6 ... 71
Topic 7. Data recognition and discriminant analysis ... 81
Laboratory work 7 ... 81
Bibliograhy ... 92
Main ... 92
Additional ... 93
Internet resources ... 94

НАВЧАЛЬНЕ ВИДАННЯ

СТАТИСТИЧНЕ МИСЛЕННЯ ДЛЯ НАУКИ ПРО ДАНІ

Методичні рекомендації до лабораторних робіт для студентів спеціальності 122 "Комп'ютерні науки" другого (магістерського) рівня

(англ. мовою)

Самостійне електронне текстове мережеве видання

Укладачі: **Раєвнєва** Олена Валентинівна
Дериховська Вікторія Ігорівна

Відповідальний за видання *О. В. Раєвнєва*

Редактор *З. В. Зобова*

Коректор *З. В. Зобова*

Подано лабораторні роботи з навчальної дисципліни та методичні рекомендації для їх виконання, що допоможуть студентам одержати практичні навички використання інструментів економіко-математичного моделювання під час вивчення складних соціально-економічних процесів та систем.

Рекомендовано для студентів спеціальності 122 "Комп'ютерні науки" другого (магістерського) рівня.

План 2020 р. Поз. № 116 ЕВ. Обсяг 96 с.

Видавець і виготовлювач – ХНЕУ ім. С. Кузнеця, 61166, м. Харків, просп. Науки, 9-А

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру
ДК № 4853 від 20.02.2015 р.