MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY
OF ECONOMICS

ECONOMETRICS

Practicum
**for Bachelor's (first) degree students
of all specialities**

Kharkiv
S. Kuznets KhNUE
2018

UDC 330.43(07.034)
E42

Compiled by: L. Guryanova
S. Prokopovych
S. Milevskiy

Затверджено на засіданні кафедри економічної кібернетики.
Протокол № 7 від 16.01.2018 р.

Самостійне електронне текстове мережеве видання

E42 **Econometrics :** practicum for Bachelor's (first) degree students of all specialities [Electronic resource] / compiled by L. Guryanova, S. Prokopovych, S. Milevskiy. – Kharkiv : S. Kuznets KhNUE, 2018. – 82 p. (English)

The basic questions of analysis and forecasting of socio-economic and financial processes and systems through the application of econometric methods and models are presented. The practicum on the academic discipline with the use of the software Microsoft Excel is provided.

For students of all specialities.

UDC 330.43(07.034)

The general information

The practicum is intended for students to assimilate the theoretical and practical material, acquire skills in the use of application packages to ensure the construction and study of different types of models, and expand students' knowledge of the application of mathematical modeling to economic calculation, prediction, and analysis of economic systems.

Microsoft Excel is proposed to be used for practical activities. This package contains a set of statistical methods that support solutions to various econometric problems. Microsoft Excel was developed to work with in Windows. The practical tasks were developed on the assumption that students are familiar with the basic principles and methods of work in Windows.

Each practical task is considered as an example for solving some problems with detailed comments and pictures. It is recommended that practical task should be performed consistently as the steps and techniques are common and will be described only once. In addition, consistent performance helps better learn the material and consolidate the knowledge of the academic discipline.

The practical tasks deal with the main topics and subjects based on the theoretical material of the relevant topics as well as previous issues. Each activity contains goals and tasks to be performed and guidelines for doing them.

To confirm the results of the practical activity students should prepare individual reports that include: the basic data for solving the problem, formulation of the problem, printing the main results of building the model, analysis of calculations and findings. The variant number, the full name of the student who performed the work and the full name of the teacher who assessed the report should be indicated on the title page.

The mark for the work depends on the practical activity results and the way it has been presented. Special attention is paid to the knowledge of the theory, correctness and completeness of the findings of the economic interpretation of the results.

Content module 1
The basics of econometric modeling

Practical activity 1. Preliminary analysis of baseline data

The goal is to consolidate the theoretical and practical material, acquire the skills in working with the Descriptive Statistics function of the Data Analysis add-in in the MS Excel package.

The task is to analyze the variation series for the sampled data of Ukrainian banks using the Descriptive Statistics function in the Data Analysis module in the MS Excel package.

1. Calculate the statistical characteristics of the series (the mean, the variance, the mean square deviation, the modal value, the median, the range of variation, the asymmetry and excess coefficients).
2. Construct a histogram and a random value distribution polygon, draw conclusions about the nature of the distribution law.
3. Test the hypothesis of a normal distribution law using the Kolmogorov – Smirnov and Pearson criteria.
4. Draw conclusions about the grouping of banks based on the size of the relevant indicator.
5. Identify abnormal observations, remove them from the sample and repeat all the previous stages of the study.
6. Compare two samples. Draw conclusions.

Guidelines

For the solution and analysis of this type of tasks, MS Excel provides the Data Analysis add-in. Let's consider the order of work in this add-in.

1.1. Launching MS Excel and data preparation. Select the MS Excel program in the application menu, after entering it, enter the baseline data as shown in Fig. 1.1.

In order to facilitate the analysis of the baseline data for their dimension in the *Income* column, select the data, the *Cell / Numerical Format* for them, and the *Bulk Batch Separator* check box (Fig. 1.2).

	A	B	C
1		**Bank**	**Income**
2	1	Privatbank	11874771
3	2	Prominvest	793821
4	3	Aval	876148
5	4	Oshchadbank	389719
6	5	Ukrsotsbank	459234
7	6	Ukrsibbank	451074
8	7	Ukreximbank	328131
9	8	Raiffeisenbank	209010
10	9	Bosom	273945
11	10	Brokbusinessbank	167741
12	11	Ukrprombank	232158
13	12	Finance and Credit	175292
14	13	First International Bank	111185
15	14	Khreshchatyk	70674
16	15	Forum	145468
17	16	Pivdennyy	132243
18	17	Pravexbank	120243
19	18	Kreditprombank	100261
20	19	UkrGasBank	104326
21	20	Credit Bank	114054
22	21	Citibank	42602
23	22	Ingbank Ukraine	35241
24	23	Vabank	71296
25	24	CreditDnipro	91436
26	25	Dongorbank	86384

Fig. 1.1. **The baseline data in the MS Excel sheet**

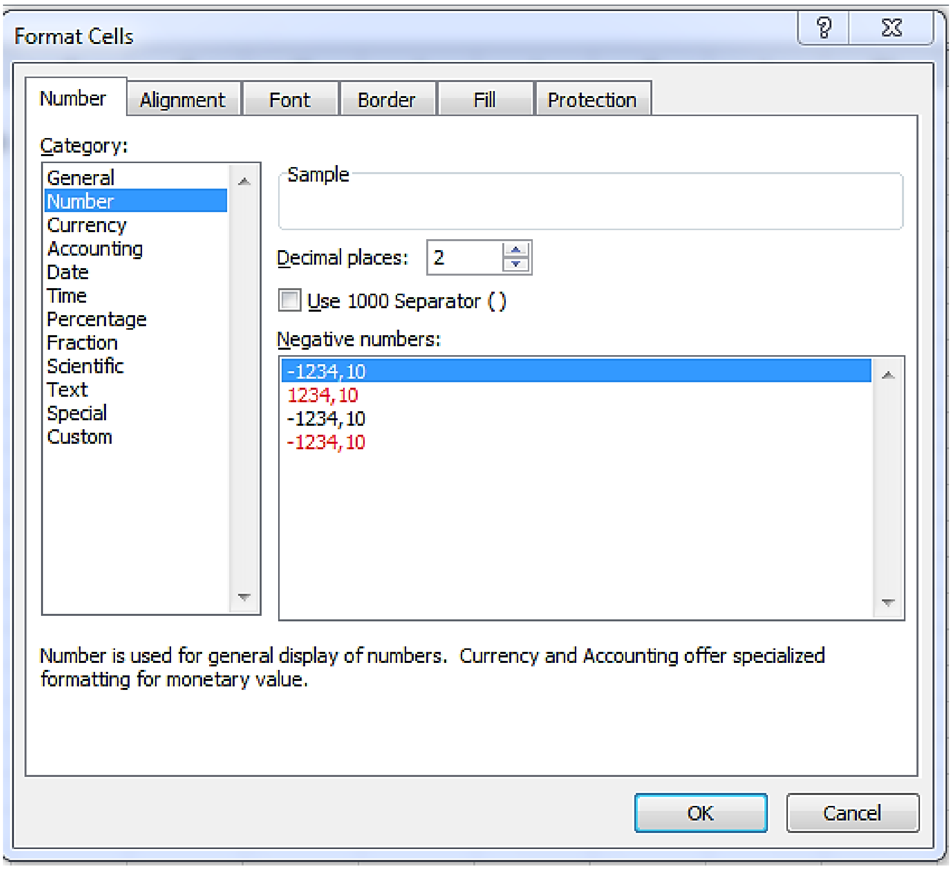


Fig. 1.2. **Setting the cell format**

After this, the initial data will look like in Fig. 1.3.

	A	B	C
1		**Bank**	**Income**
2	1	Privatbank	11 874 771
3	2	Prominvest	793 821
4	3	Aval	876 148
5	4	Oshchadbank	389 719
6	5	Ukrsotsbank	459 234
7	6	Ukrsibbank	451 074
8	7	Ukreximbank	328 131
9	8	Raiffeisenbank	209 010
10	9	Bosom	273 945
11	10	Brokbusinessbank	167 741
12	11	Ukrprombank	232 158
13	12	Finance and Credit	175 292
14	13	First International Bank	111 185
15	14	Khreshchatyk	70 674

Fig. 1.3. **The baseline data after formatting (a fragment)**

1.2. Starting Data Analysis add-in. Click the *Data* tab, click the *Data Analysis* button.

If the *Data Analysis* button is unavailable, you need to install the Add-in *Analysis Package*.

To do this, on the *File* tab, click *Parameters*, and then the *Add-ins* category. In the *Management* list, click *Excel Extras*, and then click *Go* (Fig. 1.4).

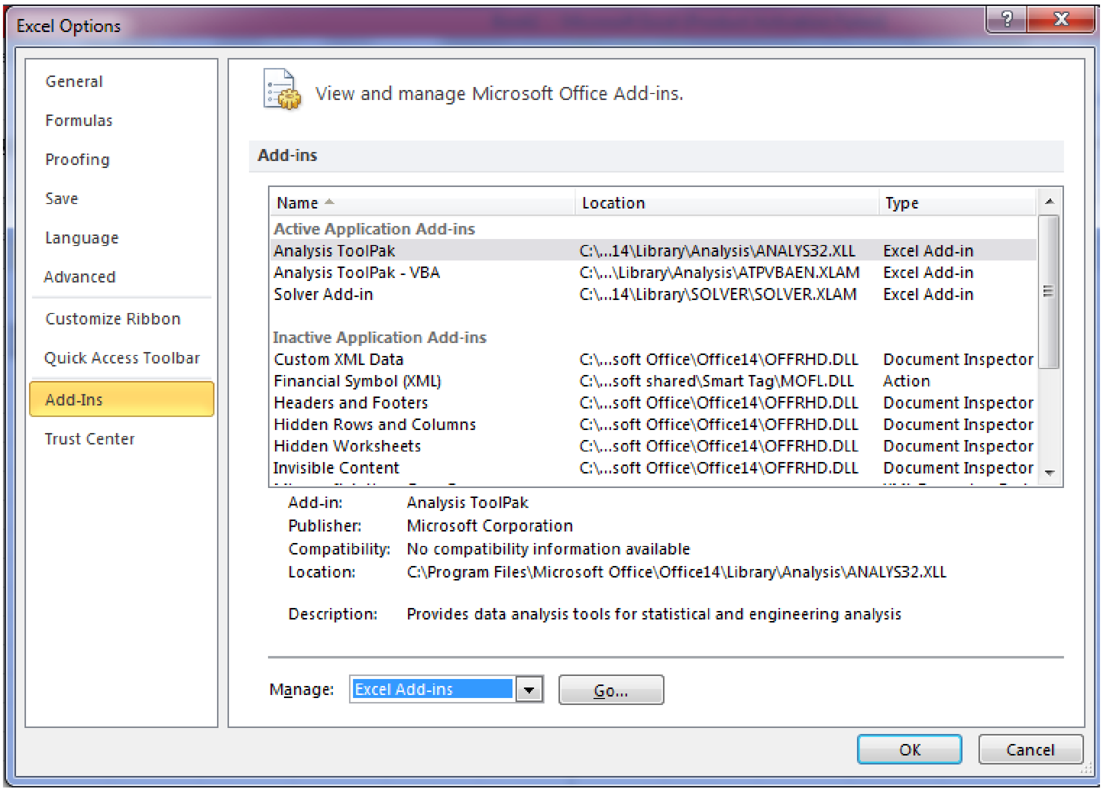


Fig. 1.4. **Add-ins Management in MS Excel**

In the *Available Add-ins* window, select the *Analysis Package* check box, and then click *OK* (Fig. 1.5).

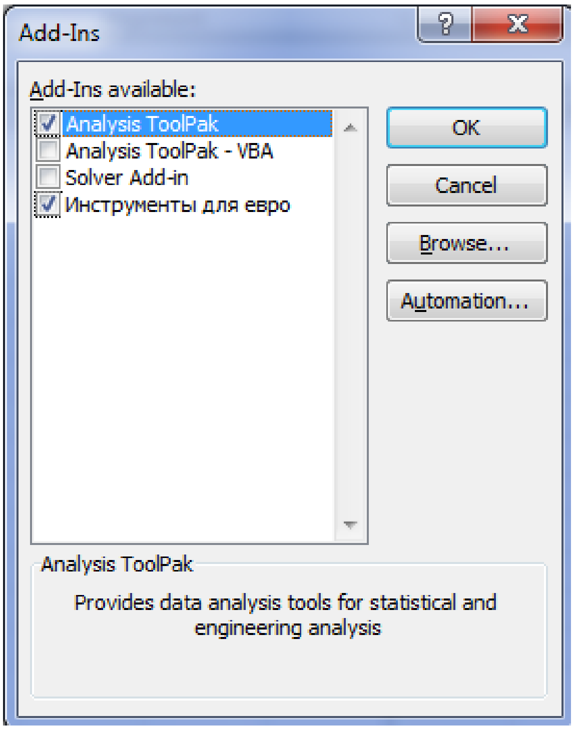


Fig. 1.5. **Selecting the Analysis Package**

If the *Analysis Package* item is not listed in the *Available Add-ins* list, click *Browse* to find the add-in.

If you receive a message that the *Analysis Package* add-in is not installed on your computer, click the *Yes* button to install it.

After clicking the *Data Analysis* button on the *Data* tab, a window with a list of available features will appear. Select *Descriptive Statistics* and click *OK* (Fig. 1.6).

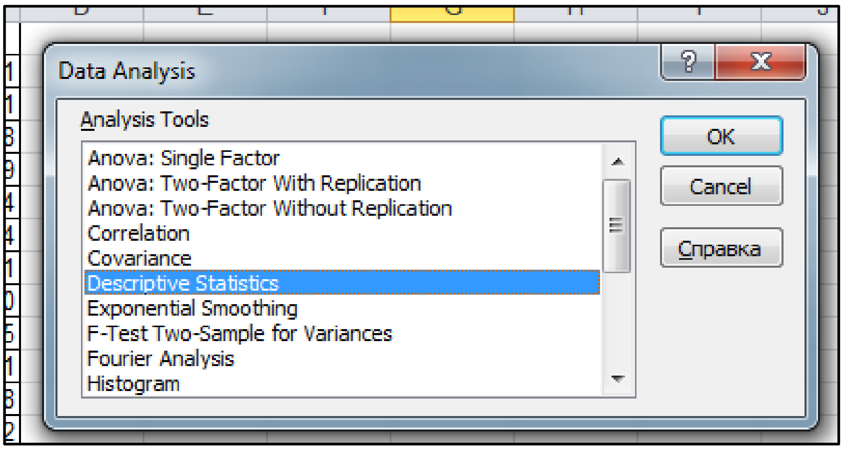


Fig. 1.6. **Selecting the function**

In the window that appears, select the output range (Fig. 1.7).

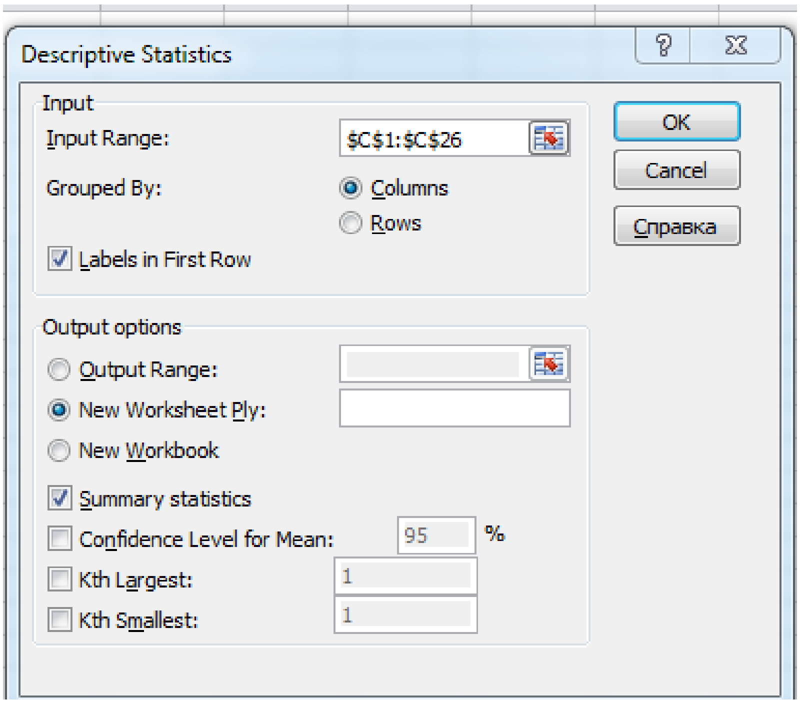


Fig. 1.7. **Selecting the source data**

Click *OK* and you will receive the results in a new sheet (Fig. 1.8).

	A	B
1	*Income*	
2		
3	Mean	698258,28
4	Standard Error	467711,871
5	Median	145468
6	Mode	#Н/Д
7	Standard Deviation	2338559,36
8	Sample Variance	5,4689E+12
9	Kurtosis	24,5130755
10	Skewness	4,93128873
11	Range	11839530
12	Minimum	35241
13	Maximum	11874771
14	Sum	17456457
15	Count	25

Fig. 1.8. **The results of the descriptive statistics calculation for a discrete series**

The following characteristics are obtained:

Mean – the mean value – a generalized indicator that characterizes the typical (averaged) values of the options for the totality of objects under investigation;

Standard Error – the standard error of the mean – the value that characterizes the standard deviation of the sample mean, calculated on the sample size *n* from the general data set. The magnitude of the standard error depends on the variance of the general data set and the sample size *n*. Since the variance of the general data set is usually unknown, the estimate of the standard error is calculated by the formula:

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}, \tag{1}$$

where *s* is standard deviation of the random variable based on the unbiased estimation of its sample dispersion;

n is the sample size;

Median – the median which is an option that characterizes the center of variation;

Mode – the modal value – which is the variant that most often occurs in the studied set of objects;

Standard Deviation – the standard deviation which is a measure of dispersion relative to the aggregate average;

Sample Variance – the unbiased sample variance estimation – an indicator that reflects the fluctuations of each option relative to its mean;

Kurtosis – the asymmetry coefficient characterizing the asymmetry of the distribution: if $K_a > 0$, then there is a right-side asymmetry, if $K_a < 0$, then it is left-side;

Skewness – the coefficient of excess characterizing the height of the distribution peak: if $K_b > 0$, there is a high vertex distribution, if $K_b < 0$, the vertex is flat;

Range – the sampling range;

Minimum – the minimum value;

Maximum – the maximum value;

Sum – the sum of all sample options;

Count – the number of observations (sample items).

Copy these characteristics to the sheet with the initial data.

2. Construction of a distribution polygon. To illustrate the representation of a discrete series, it is necessary to construct a distribution graph. The graphical representation of the discrete variation series, where the objects under investigation are located along the *OX*-axis, and the *OY*-axis is the value of the option, is called the distribution polygon (Fig. 1.9).

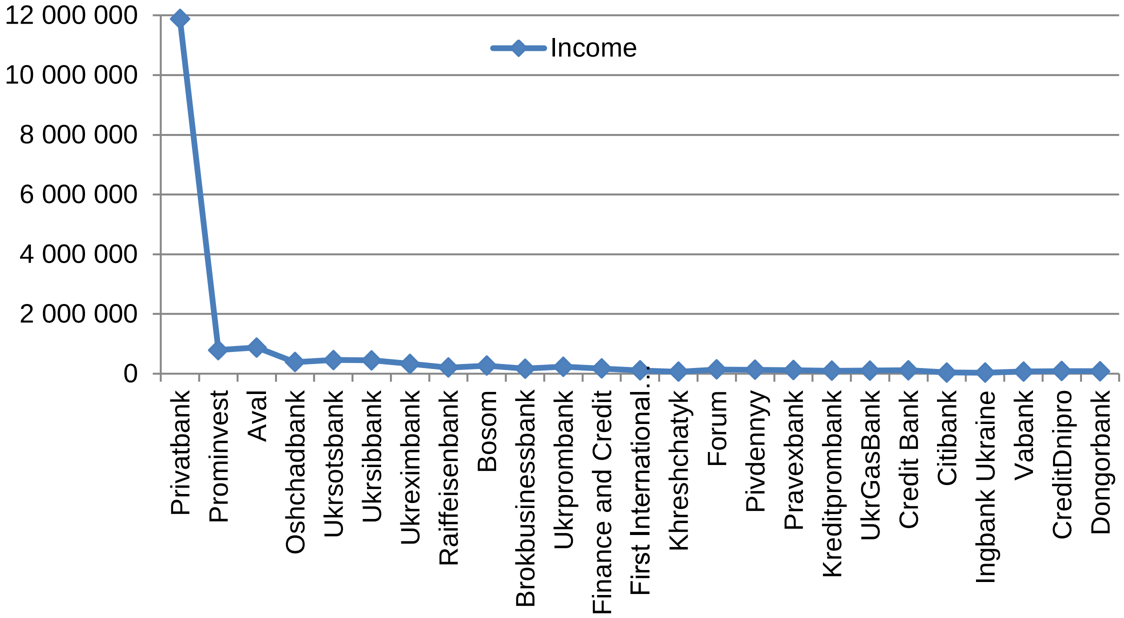


Fig. 1.9. **The distribution polygon**

3. Construction of an interval series. To transform a discrete series to an interval one, you need to calculate the step of grouping:

$$k = \frac{\max(x) - \min(x)}{1 + 3.32\log_{10} N}, \tag{2}$$

where $\max(x)$ is the maximum value of the discrete variation series;

$\min(x)$ is the minimum value of the discrete variation series;

$\max(x) - \min(x)$ is the range of the variation series;

N is the number of observations;

$1 + 3.32\log_{10} N$ is the number of intervals.

In the example that we are considering, we form $1 + 3.32\log_{10} N = 5.64 \approx 6$ intervals with the grouping step equal to $k = \dfrac{11\,839\,530}{6} = 1\,973\,255$.

The next step in the formation of an interval series is the definition of the upper and lower limits of intervals: the lower limit of the first interval is the minimum value of the variation series, the upper limit of the interval is calculated as the sum of the value of the lower limit and the step of grouping; the lower limit of the next interval is the value of the upper limit of the previous interval, etc. Calculations are carried out until the maximum value of the variation series is covered by the last interval. The formulas for calculation are shown in Fig. 1.10, and the results are presented in Fig. 1.11.

	A	B	C	D	E	F
22	21	Citibank	42602		Range	11839530
23	22	Ingbank Ukraine	35241		Minimum	35241
24	23	Vabank	71296		Maximum	11874771
25	24	CreditDnipro	91436		Sum	17456457
26	25	Dongorbank	86384		Count	25
27						
28		Number of intervals	=1+3,32*LOG10(25)			
29		*k*	=F22/6			
30						
31	*i*	lower limit of interval	upper limit of interval			
32	1	=F23	=B32+$C$29			
33	2	=C32	=B33+$C$29			
34	3	=C33	=B34+$C$29			
35	4	=C34	=B35+$C$29			
36	5	=C35	=B36+$C$29			
37	6	=C36	=B37+$C$29			

Fig. 1.10. **The formulas for calculating the interval limits**

	A	B	C	D	E	F
22	21	Citibank	42 602		Range	11 839 530
23	22	Ingbank Ukra	35 241		Minimum	35 241
24	23	Vabank	71 296		Maximum	11 874 771
25	24	CreditDnipro	91 436		Sum	17 456 457
26	25	Dongorbank	86 384		Count	25
27						
28		Number of interv	5,64116083			
29		*k*	1 973 255			
30						
31	*i*	lower limit of interval	upper limit of interval			
32	1	35241	2 008 496			
33	2	2 008 496	3 981 751			
34	3	3 981 751	5 955 006			
35	4	5 955 006	7 928 261			
36	5	7 928 261	9 901 516			
37	6	9 901 516	11 874 771			

Fig. 1.11. **The results of the calculation of the interval limits**

To complete the formation of an interval series, it is necessary to determine the frequency of occurrence of values in the corresponding interval, using the function: *FREQUENCY* (array of data, array of intervals).

To do this, you need to select the entire range of cells, which will calculate the frequencies (in this case D32: D37), write the "=" sign and enter the *FREQUENCY* () function, as shown in Fig. 1.12.

D32 | =FREQUENCY(C2:C26;C32:C37)

	A	B	C	D	E
25	24	CreditDnipro	91436		Sum
26	25	Dongorbank	86384		Count
27					
28		Number of intervals	5,641160828		
29		*k*	1973255		
30					
31	*i*	lower limit of interval	upper limit of interval	Empirical frequencies (f)	
32	1	35241	2008496	=FREQUENCY(	
33	2	2008496	3981751		
34	3	3981751	5955006		
35	4	5955006	7928261		
36	5	7928261	9901516		
37	6	9901516	11874771		

Fig. 1.12. **The formula for calculating the empirical frequencies of an interval series**

Then, at the same time, press the three Shift + Ctrl + Enter buttons. Then the whole array of frequencies will be output on a sheet. The results of calculating the empirical frequencies and some other intermediate calculations are shown in Fig. 1.13.

	A	B	C	D	E	F
31	*i*	lower limit of interval	upper limit of interval	Empirical frequencies (f)	Middle of the interval (x)	x·f
32	1	35241	2 008 496	24	1 021 868,5	24 524 844
33	2	2 008 496	3 981 751	0	2 995 123,5	0
34	3	3 981 751	5 955 006	0	4 968 378,5	0
35	4	5 955 006	7 928 261	0	6 941 633,5	0
36	5	7 928 261	9 901 516	0	8 914 888,5	0
37	6	9 901 516	11 874 771	1	10 888 143,5	10 888 144
38	**Sum**			**25**		**35 412 988**

Fig. 1.13. **The results of the interim calculations**

Below, the results of calculations of the basic statistical characteristics of the interval series are shown.

The average value:

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i} = \frac{35\,412\,988}{25} = 1146\,519.5. \quad (3)$$

In the next step, you need to perform some more intermediate calculations, the results of which are shown in Fig. 1.14.

	D	E	F	G	H
31	Empirical frequencies (f)	Middle of the interval (x)	x·f	$(x-\overline{x})^2 \cdot f$	$(x-\overline{x})^4 \cdot f$
32	24	1 021 868,5	24 524 844	3 737 985 883 224	5,82189E+23
33	0	2 995 123,5	0	0	0
34	0	4 968 378,5	0	0	0
35	0	6 941 633,5	0	0	0
36	0	8 914 888,5	0	0	0
37	1	10 888 143,5	10 888 144	89 711 661 197 376	8,04818E+27
38	**25**		**35 412 988**	**93 449 647 080 600**	**8,04876E+27**

Fig. 1.14. **The interim calculations**

Now you can calculate the following characteristics of the interval series.

The dispersion:

$$D(x) = \frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i} = \frac{93\ 449\ 647\ 080\ 600}{25} = 3\ 737\ 985\ 883\ 224. \qquad (4)$$

The standard deviation:

$$S(x) = \sqrt{D(x)} = \sqrt{3\ 737\ 985\ 883\ 224} = 1\,933\ 387.2. \qquad (5)$$

The modal value: to calculate this characteristic you need to define the modal interval. A modal interval is the interval with the highest frequency (in the example, it is the first interval). Therefore,

$$M_o = \frac{f_{mo} - f_{mo-1}}{(f_{mo} - f_{mo-1}) + (f_{mo} - f_{mo+1})} \cdot k + x_{mo} = \\ = \frac{24 - 0}{(24 - 0) + (24 - 0)} \cdot 1\,973\ 255 + 35\ 241 = 1\,021\,868.5, \qquad (6)$$

where f_{mo} is the frequency of the modal interval;

f_{mo-1} is the frequency of the interval preceding the modal;

f_{mo+1} is the frequency of the interval following the modal;

k is the intervals' grouping step;

x_{mo} is the lower bound of the modal interval.

The median: to calculate this characteristic, it is necessary to determine the median interval. The median interval is the interval that covers the half-sum of all frequencies, that is to determine the median, it is necessary to calculate the half-sum of all frequencies $N_{me} = \frac{\sum f_i}{2} = \frac{25}{2} = 12.5$ and the accumulated frequencies of each interval (Fig. 1.15).

	D	I
31	Empirical frequencies (f)	Cumulative frequencies (S)
32	24	24
33	0	24
34	0	24
35	0	24
36	0	24
37	1	25
38	25	

Fig. 1.15. **Calculation of the accumulated frequencies**

In the example, half-sum of all the frequencies covers the first interval. Therefore,

$$M_e = \frac{N_{me} - S_{me-1}}{f_{me}} \cdot k + x_{me} = \qquad (7)$$

$$= \frac{12.5 - 0}{24} \cdot 1\,973\,255 + 25\,241 = 1\,062\,978.0,$$

where f_{me} is the frequency of the median interval;

S_{me-1} is the cumulative (accumulated) frequency of the interval preceding the median;

N_{me} is the half-sum of all the frequencies;

k is the intervals' grouping step;

x_{me} is the lower boundary of the median interval.

The asymmetry coefficient:

$$K_a = \frac{\bar{x} - M_o}{S(x)} = 0.204. \qquad (8)$$

Because, $K_a > 0$, there is a right-side asymmetry.

The coefficient of excess (for interim calculations, see Fig. 1.14):

$$K_e = \frac{1}{\sum f_i} \cdot \frac{\sum (x_i - \bar{x})^4 f_i}{(S(x))^4} = 23.042. \qquad (9)$$

Because $K_e > 0$, we conclude that there is a high vertex distribution.

The results of the calculation of the interval series characteristics are shown in Fig. 1.16.

	A	B	C
40	**Interval series**		
41		Mean	1 416 519,5
42		Sample Variance	3 737 985 883 224,0
43		Standard Deviation	1 933 387,2
44		Mode	1 021 868,5
45		Median	1 062 978,0
46		Kurtosis	0,204124145
47		Skewness	23,042

Fig. 1.16. **The results of the calculations of the interval series characteristics**

The graphic representation of the interval series is the distribution histogram (Fig. 1.17).

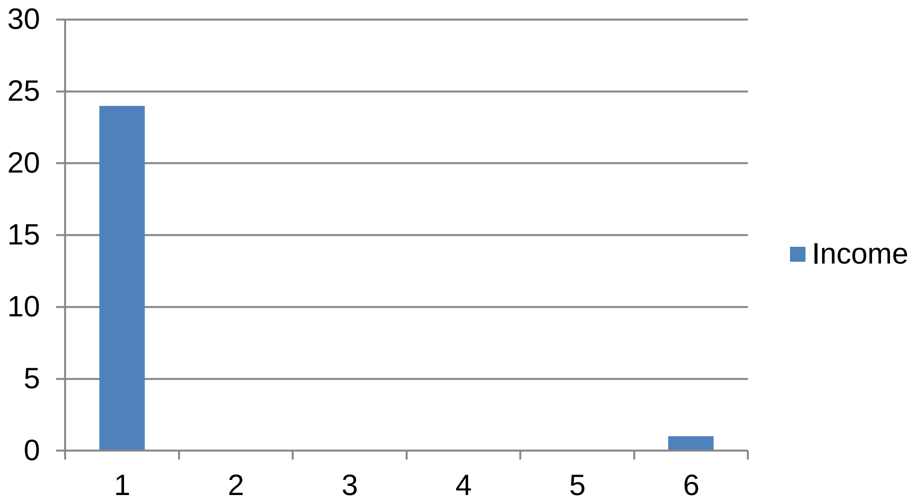


Fig. 1.17. **The distribution histogram**

4. Checking the sample for the normal distribution law. Further analysis involves checking the sample for the normal distribution law based on the Pearson and Kolmogorov – Smirnov criteria.

A prerequisite for determining the nature of the distribution law is the calculation of theoretical frequencies corresponding to the normal distribution law:

$$f'(t) = \frac{n \cdot k}{S(x)} \cdot \varphi(t), \qquad (10)$$

$$\varphi(t) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}, \tag{11}$$

$$t = \frac{x - \bar{x}}{S(x)}, \tag{12}$$

where $f'(t)$ is the theoretical frequencies corresponding to the normal distribution law,

x_i is the value of the i variable;

$\bar{x}$ is the average value;

$S(x)$ is the standard deviation;

n is the number of observations;

k is the grouping step.

For further calculations, you need to create a new table with empirical frequencies and mid-intervals. To determine the Gauss function $\varphi(t)$, you need to use the formula *NORM.DIST (x; mean; standard_dev; integral)*, which returns the normal distribution function for the specified mean and standard deviation. So, for the first interval, in the cell D50, enter the formula =1-*NORM.DIST(C50;$C$41;$C$43;1)*, as shown in Fig. 1.18.

40		Mean	=F38/D38			
41		Sample Variance	=G38/D38			
42		Standard Deviation	=SQRT(C42)			
43		Mode	=(D32-0)/((D32-0)+			
44		Median	=12,5/24*C29+B3:			
45		Kurtosis	=(C41-C44)/C43			
46		Skewness	=1/25*H38/C43^4			
47						
48	*i*	Empirical frequencies (f)	Middle of the interval (x)	φ(t)	Theoretical frequencies (f')	Cumul. Theoret. Frequencies (S')
49	1	=D32	=E32	=1-NORM.DIST(C50;C	=CEILING(25/C43*D50*(	=E50
50	2	=D33	=E33	=1-NORM.DIST(C51;C	=CEILING(25/C43*D50*(	=F50+E51
51	3	=D34	=E34	=1-NORM.DIST(C52;C	=CEILING(25/C43*D50*(	=F51+E52
52	4	=D35	=E35	=1-NORM.DIST(C53;C	=CEILING(25/C43*D50*(	=F52+E53
53	5	=D36	=E36	=1-NORM.DIST(C54;C	=CEILING(25/C43*D50*(	=F53+E54
54	6	=D37	=E37	=1-NORM.DIST(C55;C	=CEILING(25/C48*D50*(	=F54+E55
55	**Sum**	**25**				

Fig. 1.18. **The formulas for calculating the theoretical and accumulated theoretical frequencies**

The results of calculations are presented in Fig. 1.19.

	A	B	C	D	E	F
49	*i*	Empirical frequencies (f)	Middle of the interval (x)	φ(t)	Theoretical frequencies (f')	Cumul. Theoret. Frequencies (S')
50	1	24	1 021 868,5	0,58087	15	15
51	2	0	2 995 123,5	0,20711	6	21
52	3	0	4 968 378,5	0,03310	1	22
53	4	0	6 941 633,5	0,00213	1	23
54	5	0	8 914 888,5	0,00005	1	24
55	6	1	10 888 143,5	0,00000	1	25
56	**Sum**	**25**			**25**	

Fig. 1.19. **The results of the calculation of the theoretical and accumulated theoretical frequencies**

To determine the nature of the distribution law and its compliance with the normal law, the Pearson criterion is most often used, calculated by the formula:

$$\chi^2 = \sum \frac{(f - f')^2}{f'}, \tag{13}$$

where f is the empirical frequencies;

f' is the theoretical frequencies.

The estimated value of the criterion must be compared with the critical one, which is in a special table and depends on the accepted probability and the number of degrees of freedom ($k = m - 3$, where m is the number of intervals). If $\chi^2 \leq \chi^2_{tab}$, then the difference between the empirical and theoretical frequencies can be considered random and the hypothesis regarding the normal distribution law cannot be rejected.

The Kolmogorov – Smirnov criterion is used for determining the maximum difference between the frequencies of the empirical and theoretical distribution:

$$\lambda = \frac{D}{\sqrt{\sum f}}, \tag{14}$$

where $D = |S - S'|$ is the maximum difference between the accumulated empirical and theoretical frequencies,

$\sum f$ is the sum of empirical frequencies.

The estimated value of the criterion is used to find the probability of adopting a hypothesis regarding the normal distribution law. The greater the probability value, the greater the probability that the discrepancies between the empirical and theoretical frequencies are random.

Fig. 1.20 shows the results of the intermediate calculations and the estimated values of the Pearson and Kolmogorov – Smirnov criteria.

Empirical frequencies (f)	Middle of the interval (x)	φ(t)	Theoretical frequencies (f')	Cumul. Theoret. Frequencies (S')	$(f - f')^2/f'$	Cumulative frequencies (S)	|S - S'|
24	1 021 868,5	0,58087	15	15	5,4	24	9
0	2 995 123,5	0,20711	6	21	6	24	3
0	4 968 378,5	0,03310	1	22	1	24	2
0	6 941 633,5	0,00213	1	23	1	24	1
0	8 914 888,5	0,00005	1	24	1	24	0
1	10 888 143,5	0,00000	1	25	0	25	0
25			**25**		**14,4**	**Max =**	**9**

Criterion	Estimated value	Critical Criteria Values (α=0,05)
Pearson	14,4	7,81
Kolmogorov-Smirnov	1,8	0,27

Fig. 1.20. **The calculated values of the Pearson and Kolmogorov – Smirnov criteria**

Comparing the obtained values with the tables, we can conclude that the differences between the empirical and theoretical distribution are significant, so the hypothesis regarding the normal law of the distribution of the random variable should be discarded.

Thus, this sample is not homogeneous and close to normal distribution. This means that it cannot be used as a source for building an econometric model, and it needs to be refined.

5. Removal of abnormal observations. An anomalous observation, that is, such that is not typical of this sample, can be detected by means of analysis of the distribution polygon (see Fig. 1.9) or distribution histogram (see Fig. 1.17). There is at least one such observation – Privatbank. It has an income that exceeds dozens of times all other observations and should be excluded from the sample.

Copy all data except Privatbank to a new sheet and repeat all the previous steps of the study.

Compare the characteristics of both samples, draw conclusions.

Practical activity 2. Modeling and analysis of simple linear econometric models

The goal is to consolidate the theoretical and practical material, to acquire the skills in modeling and analysis of simple econometric models in the Data Analysis add-in of Microsoft Excel.

The task is to verify the existence of linear dependence between the given indicators in the Data Analysis add-in of Microsoft Excel.

1. Build a linear econometric model and define all its specifications (the parameters of the model, the standard deviations of the parameters of the model, the variance and standard deviation of the error of the model, the correlation and determination coefficients).

2. Verify the statistical significance of parameters and the correlation coefficient using the Student's t-test. Verify the adequacy of the model with the Fisher test.

3. Calculate theoretical values of the dependent variable and errors of the model, construct a graph of the linear function with confidence intervals, construct a histogram and a graph of the distribution of errors, grouping of data depending on the values of errors, and provide economic interpretation of the grouping.

4. Calculate the predicted values of the dependent variable and the confidence intervals if the values of the independent parameter are provided.

5. Draw conclusions about the adequacy of the constructed model, provide economic interpretation of the dependence and the possibility of its theoretical use.

Guidelines

For the construction and analysis of simple linear econometric models, the Data Analysis add-in is provided in Microsoft Excel. Let us examine the order of work in the module.

1. Launching MS Excel and data preparation.

After starting Microsoft Excel, the input data of the model should be entered, as shown in Fig. 2.1.

	A	B	C	D	E
1	i	X_i	Y_i		
2	1	73	0,5		
3	2	85	0,7		
4	3	102	0,9		
5	4	115	1,1		
6	5	122	1,4		
7	6	126	1,4		
8	7	134	1,7		
9	8	147	1,9		
10					
11					

Fig. 2.1. **The input data**

2. Calculations.

In order to start calculations, it is necessary to select the menu item *Data / Data Analysis* (Fig. 2.2). After selection of the *Regression* tool of the module, a dialog box of the module will appear where variables for analysis can be set (Fig. 2.3).

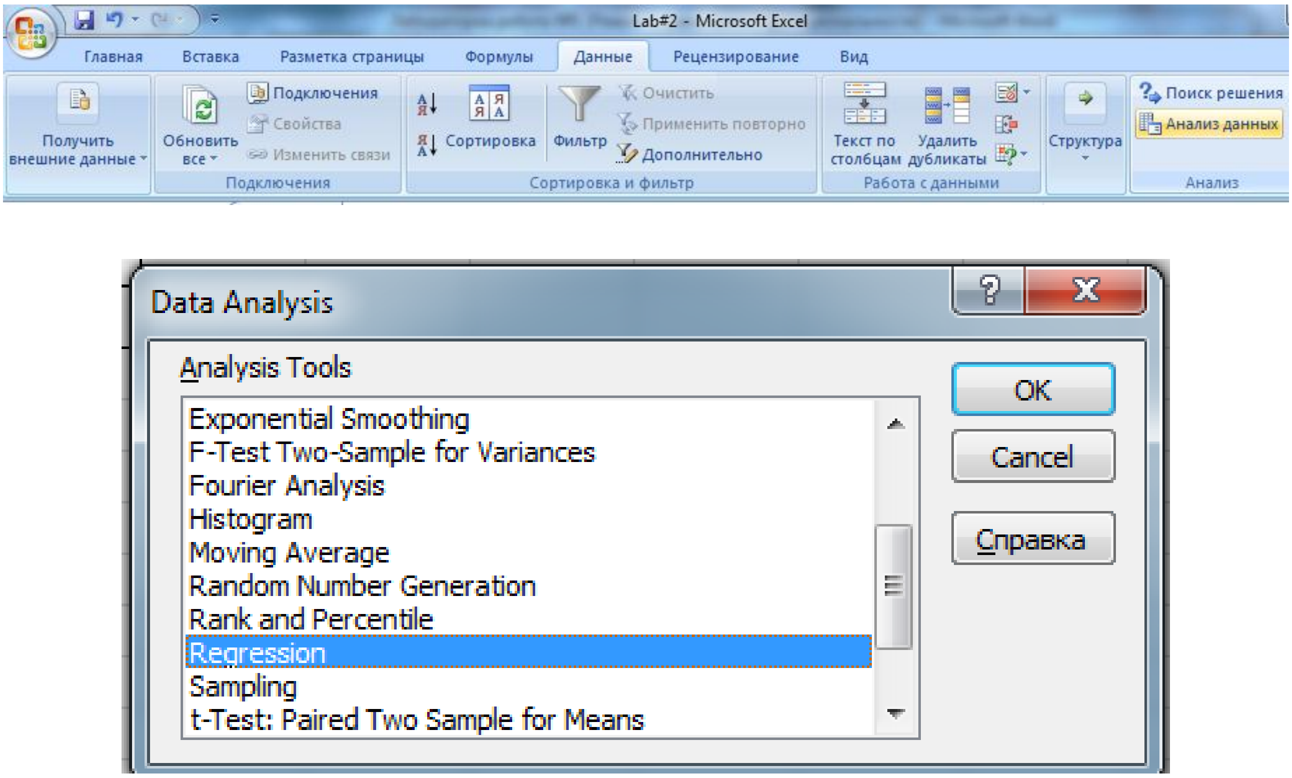


Fig. 2.2. **Selecting the module**

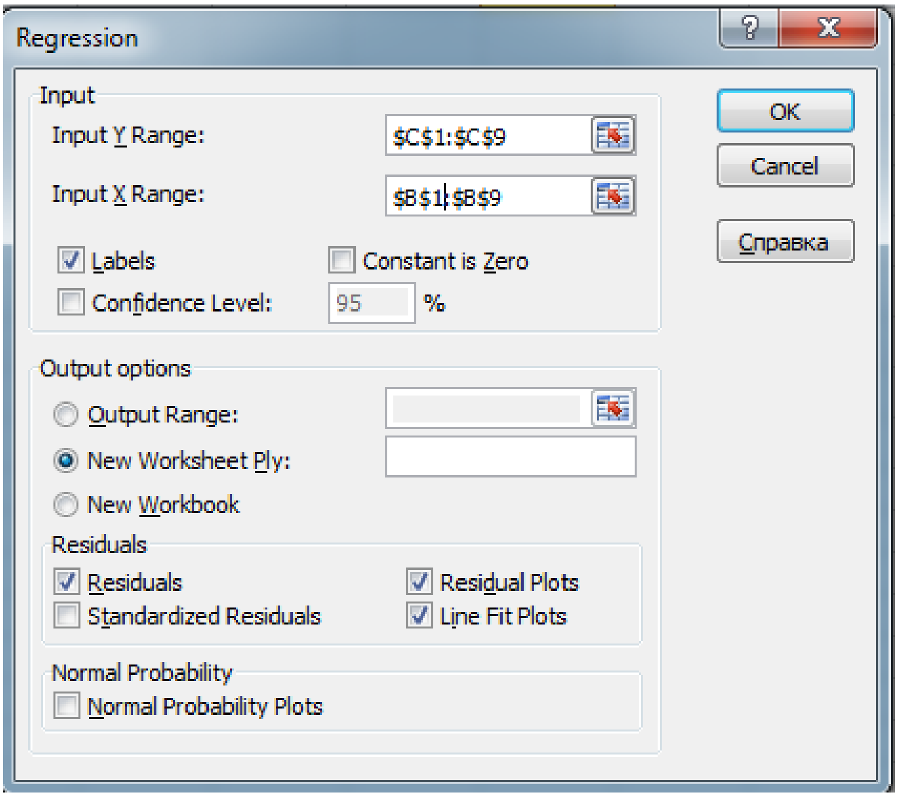


Fig. 2.3. **The dialog box of the regression tool**

Select the range of input variables *X* and *Y* (if the names of the variables are required with their initial values, the *Labels* option must be selected). If a_0 equals zero, the *Constant is Zero* option must be set. In the module, there is also a possibility to change the confidence level of the model with the *Confidence Level* option that by default equals 95 %.

In the next block of the dialog box, the output destination can be chosen. There are three options: 1) the range selected by the user (*Output Range*); 2) a new worksheet of the same file (*New Worksheet Ply*); 3) a new MS Excel file (*New Workbook*).

In the last block of the dialog box, the *Residuals* and *Residual Plots* options, that are necessary for analysis, must be chosen.

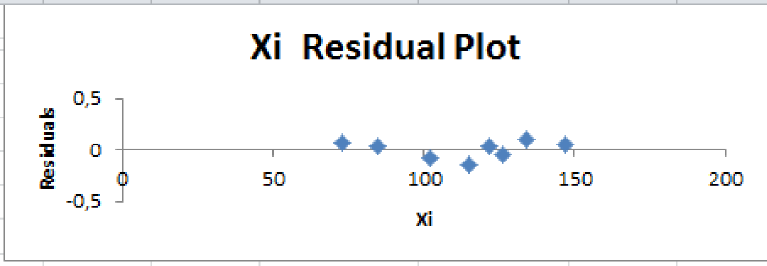
Once the selection is confirmed by pressing *OK*, the results of modeling will appear in the new worksheet of the same file (Fig. 2.4).

3. Analysis of the model, definition of its specifications, verification of the adequacy and statistical significance of the model.

Developing a linear econometric model and defining all of its specifications. The parameters received during estimation of the model are given in the third table from the top (Fig. 2.5).

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,985475971
R Square	0,971162888
Adjusted R Square	0,966356703
Standard Error	0,089321148
Observations	8



ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1,61213	1,61213	202,0652	7,57632E-06
Residual	6	0,04787	0,007978		
Total	7	1,66			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95,0%*	*pper 95,0%*
Intercept	-0,973875115	0,156155	-6,23658	0,000787	-1,355973153	-0,591777077	-1,35597	-0,59178
Xi	0,019237833	0,001353	14,21496	7,58E-06	0,015926303	0,022549363	0,015926	0,022549

RESIDUAL OUTPUT

Observation	*Predicted Yi*	*Residuals*
1	0,430486685	0,069513
2	0,66134068	0,038659
3	0,988383838	-0,08838
4	1,238475666	-0,13848
5	1,373140496	0,02686
6	1,450091827	-0,05009
7	1,60399449	0,096006
8	1,854086318	0,045914

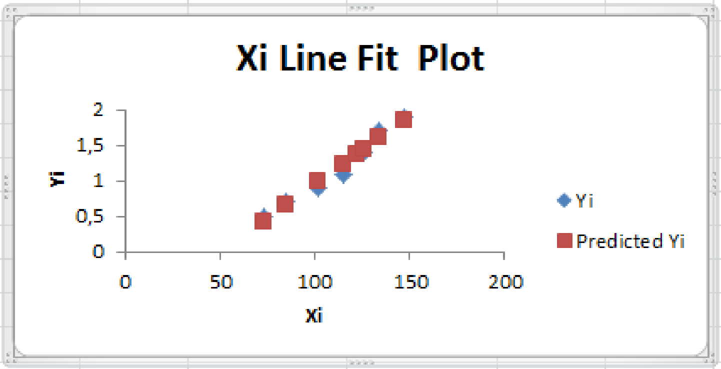


Fig. 2.4. **The modeling results**

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-0,973875115	0,156155	-6,23658	0,000787	-1,355973153	-0,59178
X Variable 1	0,019237833	0,001353	14,21496	7,58E-06	0,015926303	0,022549

Fig. 2.5. **The results (parameters) of the model estimation**

In the first column of the table (Fig. 2.5), model parameters a_0 (*Intercept*) and a_1 (X_1) are given. Thus, the theoretic model is as follows:

$$Y = -0.973875115 + 0.019237833 \cdot X_1. \qquad (15)$$

The next columns relate to the analysis of the statistical significance of the model parameters, namely, *Std Error* refers to σ_{a0} and σ_{a1} respectively; *t Stat* and *P-values* are the corresponding values of the Student's t-test for each parameter and the probability level of the error of accepting the hypothesis. The values of the latter are as follows t_{a0} = –6.236584132 and t_{a1} = 14.2149647. Both calculated values of the Student criterion exceed the tabulated one t_{tabl} = 2.447 in their absolute values, which indicates the statistical significance of both parameters (the tabulated value of this criterion can be obtained with the TINV (0.05, 6) function, where the first number is the probability of the error level, and the second one is the number of degrees of freedom).

The last two columns contain the values of the interval estimates of each model parameter with the probability level of 95 %.

Let us analyze the results of the adequacy analysis of the overall model presented in the first table of the results (Fig. 2.6).

Regression Statistics	
Multiple R	0.985475971
R Square	0.971162888
Adjusted R Square	0.966356703
Standard Error	0.089321148
Observations	8

Fig. 2.6. **The adequacy analysis of the overall model as part of the results of the model estimation**

Multiple R is the multiple correlation coefficient (in the case of simple linear regression it equals the bivariate correlation coefficient between *X* and *Y*);

R Square is the coefficient of determination of the model;

Adjusted R Square is the coefficient of determination adjusted by the number of observations and the number of model parameters;

Standard Error is the standard deviation of the model errors; this statistic is a measure of scatter of the values relative to the regression line (σ_e);

Observations correspond to the number of initial observations.

The results of the analysis of variance of the model are shown in the second table of the regression results (Fig. 2.7).

ANOVA	*df*	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.612130395	1.612130395	202.0652216	7.57632E-06
Residual	6	0.047869605	0.007978268		
Total	7	1.66			

Fig. 2.7. **The analysis of variance**

The table contains the sum of squares (*SS*) and variance (*MS*) for regression and for errors, and the Fisher's test.

The calculated value of the Fisher's test significantly exceeds its tabulated values F_{tabl} (0.05, 1, 6) = 5.987, indicating the statistical significance of the overall model (the tabulated value of this test can be obtained with the FINV(0.05, 1, 6) function, where the first number is the level of probability of error, and the last two are the numbers of degrees of freedom).

Thus, the values of these coefficients indicate a rather high level of quality and adequacy of the model, which makes it possible to use the model for forecasting.

4. Analysis of errors.

The theoretical values of the dependent variable and the error of the model are shown in the last table of the results (Fig. 2.8).

Observation	*Predicted Y*	*Residuals*
1	0.430486685	0.069513315
2	0.66134068	0.03865932
3	0.988383838	-0.088383838
4	1.238475666	-0.138475666
5	1.373140496	0.026859504
6	1.450091827	-0.050091827
7	1.60399449	0.09600551
8	1.854086318	0.045913682

Fig. 2.8. **The analysis of errors of the model**

The scatter plot of the model errors is depicted in Fig. 2.9.

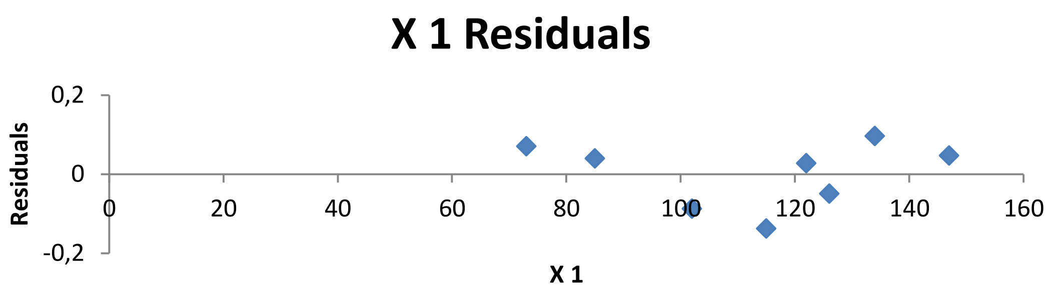


Fig. 2.9. **The residuals scatter plot of the model**

For visual verification of frequency distribution of errors by the normal probability law, errors should be grouped in advance, i.e. their variation distribution should be transferred to the normal probability plot (*see the Guidelines for laboratory work 1*).

The intermediate calculations and the resulting interval series of errors are represented in Fig. 2.10 and 2.11.

22	CONCLUSION READY			
23				
24	*Observation*	*Predicted Y*	*Remains*	
25	1	0,43048668503214	0,0695133149678604	
26	2	0,661340679522498	0,0386593204775023	
27	3	0,988383838383838	-0,088383838383838	
28	4	1,23847566574839	-0,138475665748393	
29	5	1,37314049586777	0,02685950412231	
30	6	1,45009182736455	-0,050091827365545	
31	7	1,60399449035813	0,0960055096418735	
32	8	1,85408631772268	0,045913682277319	
33				
34	max	=MAX(C25:C32)		
35	min	=MIN(C25:C32)		
36	swing	=B35-B36		
37	number of intervals	=ROUND(1+3,32*LOG10(8);0)		
38	step	=B37/B38		
39				
40	number of intervals	Lower boundary	Upper boundary	Frequency f
41	1	=B36	=B42+$B$39	=FREQUENCY($C$25:$C$32;$C$42:$C$45)
42	2	=C42	=B43+$B$39	=FREQUENCY($C$25:$C$32;$C$42:$C$45)
43	3	=C43	=B44+$B$39	=FREQUENCY($C$25:$C$32;$C$42:$C$45)
44	4	=C44	=B45+$B$39	=FREQUENCY($C$25:$C$32;$C$42:$C$45)

Fig. 2.10. **Grouping of errors. The formulas for calculations**

22	CONCLUSION READY			
23				
24	*Observation*	*Predicted Y*	*Remains*	
25	1	0,43048668503214	0,06951331496786	
26	2	0,66134067952250	0,03865932047750	
27	3	0,98838383838384	-0,08838383838384	
28	4	1,23847566574839	-0,13847566574839	
29	5	1,37314049586777	0,02685950412231	
30	6	1,45009182736455	-0,05009182736555	
31	7	1,60399449035813	0,09600550964187	
32	8	1,85408631772268	0,04591368227732	
33				
34	max	0,09600551		
35	min	-0,138475666		
36	swing	0,234481175		
37	number of intervals	4		
38	step	0,058620294		
39				
40	number of intervals	Lower boundary	Upper boundary	Frequency f
41	1	-0,138475666	-0,079855372	2
42	2	-0,079855372	-0,021235078	1
43	3	-0,021235078	0,037385216	1
44	4	0,037385216	0,09600551	4

Fig. 2.11. **Grouping of errors. The formulas of calculations**

The histogram in Fig. 2.12 represents the frequency distribution of errors.

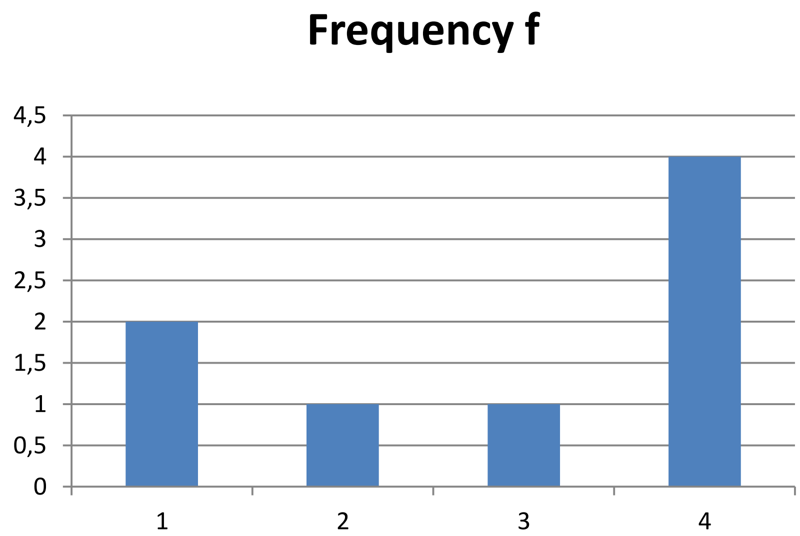


Fig. 2.12. **The histogram of the frequency distribution of errors**

Visual analysis of the histogram indicates the deviation of the distribution of errors from the normal distribution law.

5. The graph of a linear function with confidence intervals.

In order to construct a linear function with confidence intervals, first of all, it is necessary to move the theoretical values of *Y*, the value of the mean square error σ_e, the table value of Student's statistics from the page with the results to the page with the input data and calculate the average value of *X*, as shown in Fig. 2.13.

	A	B	C	D
1	i	X_i	Y_i	*Predicted* Y_i
2	1	73	0,5	0,4305
3	2	85	0,7	0,6613
4	3	102	0,9	0,9884
5	4	115	1,1	1,2385
6	5	122	1,4	1,3731
7	6	126	1,4	1,4501
8	7	134	1,7	1,6040
9	8	147	1,9	1,8541
10				
11	$\bar{X}$	113	σ_e	0,089321
12			t table	2,447

Fig. 2.13. **The initial view of the page with the input data**

The next step is to calculate the value ΔY for each of the values of the independent variable X by the formula:

$$\Delta Y_{pr} = t_{tabl} \cdot \sigma_e \cdot \sqrt{\frac{1}{n} + \frac{(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}}, \tag{16}$$

and identify the lower $(\hat{Y}_i - \Delta Y_{pr})$ and upper $(\hat{Y}_i + \Delta Y_{pr})$ boundaries of the interval series for the theoretical values of Y. The formulas for calculation are given in Fig. 2.14.

	A	B	C	D	E	F	G	H
1	i	X_i	Y_i	Prediction Y_i	ΔY	Y_i - ΔY_i	Y_i - ΔY_i	$(X_i - \overline{X})^2$
2	1	73	0,5	0,4305	`=$D$12*$D$11*SQRT(1/8+(B2-113)^2/2/$H$10)`			`=(B2-$B$11)^2`
3	2	85	0,7	0,6613	`=$D$12*$D$11*SQRT(1/8+(B3-113)^2/2/$H$10)`			`=(B3-$B$11)^2`
4	3	102	0,9	0,9884	`=$D$12*$D$11*SQRT(1/8+(B4-113)^2/2/$H$10)`			`=(B4-$B$11)^2`
5	4	115	1,1	1,2385	`=$D$12*$D$11*SQRT(1/8+(B5-113)^2/2/$H$10)`			`=(B5-$B$11)^2`
6	5	122	1,4	1,3731	`=$D$12*$D$11*SQRT(1/8+(B6-113)^2/2/$H$10)`			`=(B6-$B$11)^2`
7	6	126	1,4	1,4501	`=$D$12*$D$11*SQRT(1/8+(B7-113)^2/2/$H$10)`			`=(B7-$B$11)^2`
8	7	134	1,7	1,604	`=$D$12*$D$11*SQRT(1/8+(B8-113)^2/2/$H$10)`			`=(B8-$B$11)^2`
9	8	147	1,9	1,8541	`=$D$12*$D$11*SQRT(1/8+(B9-113)^2/2/$H$10)`			`=(B9-$B$11)^2`
10								
11	$\overline{X}$	113	σ_e	0,089321			$\sum$	`=SUM (H2:H9)`
12			t tabl	2,447				

Fig. 2.14. **The formulas for calculation of the interval values of *Y***

The calculation results are shown in Fig. 2.15.

	A	B	C	D	E	F	G	H
1	i	X_i	Y_i	Predicted Y_i	ΔY_i	Y_i - ΔY_i	Y_i + ΔY_i	$(X_i - \overline{X})^2$
2	1	73	0,5	0,4305	0,153	0,347	0,653	1600
3	2	85	0,7	0,6613	0,121	0,579	0,821	784
4	3	102	0,9	0,9884	0,085	0,815	0,985	121
5	4	115	1,1	1,2385	0,078	1,022	1,178	4
6	5	122	1,4	1,3731	0,083	1,317	1,483	81
7	6	126	1,4	1,4501	0,088	1,312	1,488	169
8	7	134	1,7	1,6040	0,104	1,596	1,804	441
9	8	147	1,9	1,8541	0,137	1,763	2,037	1156
10							Σ	4356
11	$\overline{X}$	113	σ_e	0,089321				
12			t table	2,447				

Fig. 2.15. **The results of the calculation of the interval values of *Y***

Getting all the necessary results makes it possible to start constructing the graph. After selecting the desired type of graphics *Spot*, we turn to the definition of the data series, including: the actual value of *Y*, the theoretical (predictive) value of *Y*, the lower boundary of the theoretical (predictive) value of *Y*, the upper boundary of the theoretical (predictive) value of *Y*.

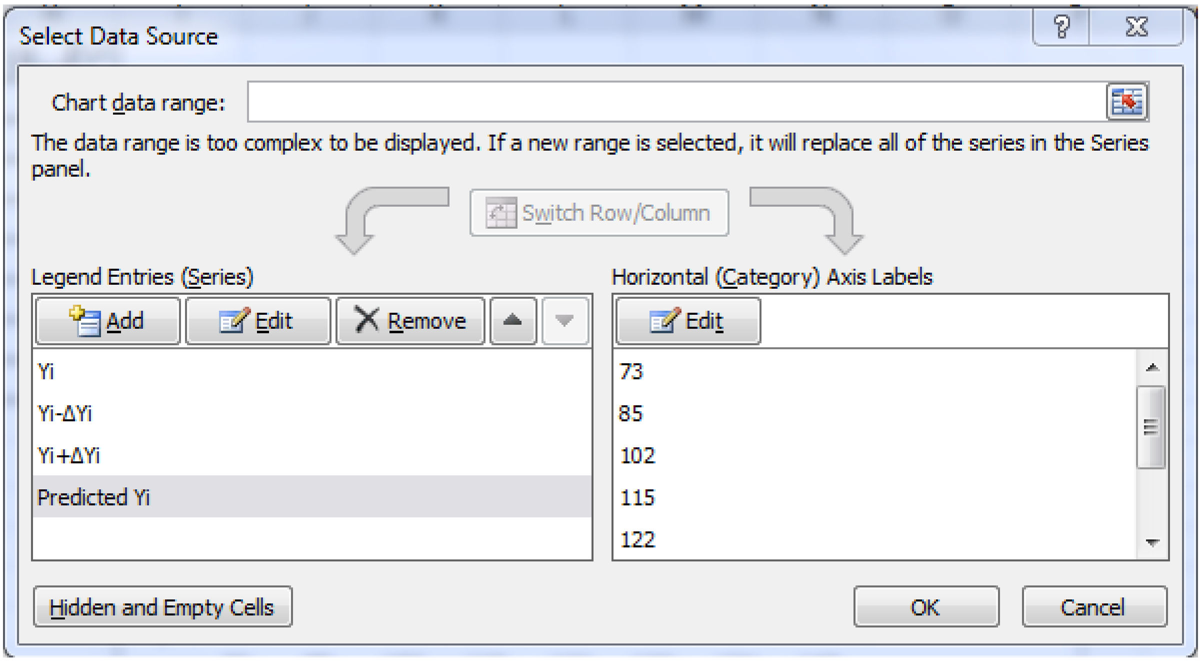


Fig. 2.16. **Selection of the data series for plotting**

The values of *X* should be chosen as the values of the horizontal axis (Fig. 2.17).

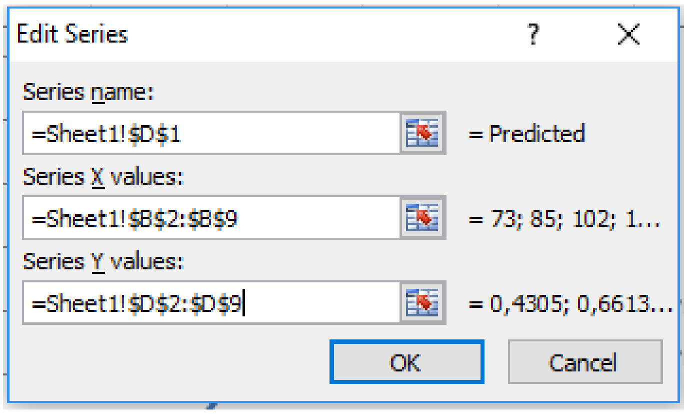


Fig. 2.17. **Selection of the initial data for plotting**

After the final formatting of some series of data (namely: the image graphs of the theoretical values of *Y* according to interval estimates dashed line through the *Data series format / Line type*) we obtain a graph of the linear function with confidence intervals (Fig. 2.18)

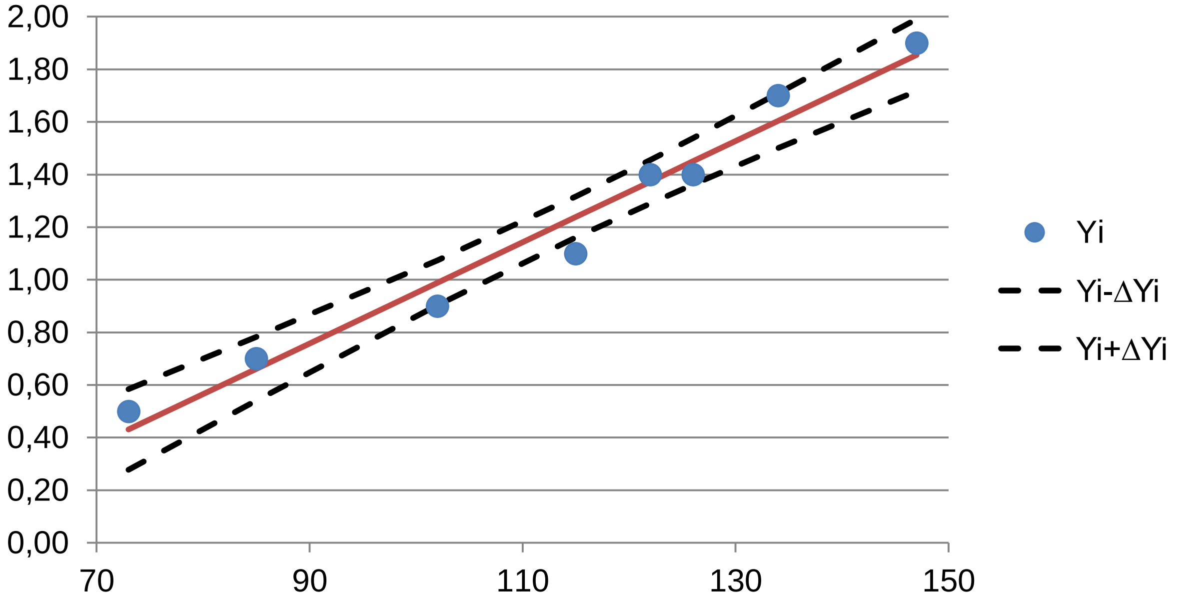


Fig. 2.18. **The graph of the linear function with confidence intervals**

As shown in Fig. 2.18 almost all the actual values of Y lie on the line corresponding to the theoretical model, confirming its high quality.

6. Calculation of predicted values of the dependent variable and confidence intervals of change.

As the model is adequate, its parameters are significant, the model can be used for forecasting. In order to calculate the predicted values of the dependent variable, it is necessary to add an additional line with the predicted values of *X* to the initial data, and calculate the theoretical (predicted) point value of *Y* using the model with point values of parameters, Δ*Ypr* by the formula (16), and the interval predicted value (Fig. 2.19).

	A	B	C	D	E	F	G	H
1	i	X_i	Y_i	*Predicted* Y_i	ΔY_i	$Y_i - \Delta Y_i$	$Y_i + \Delta Y_i$	$(X_i - \overline{X})^2$
2	1	73	0,5	0,4305	0,153	0,347	0,653	1600
3	2	85	0,7	0,6613	0,121	0,579	0,821	784
4	3	102	0,9	0,9884	0,085	0,815	0,985	121
5	4	115	1,1	1,2385	0,078	1,022	1,178	4
6	5	122	1,4	1,3731	0,083	1,317	1,483	81
7	6	126	1,4	1,4501	0,088	1,312	1,488	169
8	7	134	1,7	1,6040	0,104	1,596	1,804	441
9	8	147	1,9	1,8541	0,137	1,763	2,037	1156
10	FORECAST	140		1,7194	0,118	1,601	1,838	
11		*Coefficients*						
12	Intercept	-0,97388						
13	X_i	0,01924					Σ	4356

Fig. 2.19. **Calculation of the predictive value of *Y* with confidence intervals**

Thus, if X_{pr} = 140, we get:

$$1.601 \leq Y_{pr} \leq 1.838. \tag{17}$$

Practical activity 3. Building and analysis of multiple linear econometric models. Multicollinearity

The goal is to consolidate the theoretical and practical material on the topics "Multiple linear regression" and "Multicollinearity", to acquire the skills in modeling and analysis of multifactorial econometric models in Microsoft Excel.

The task is to verify the existence of linear multiple connection between the coincident indicators in the Data Analysis add-in of Microsoft Excel.

1. Build a linear multifactorial econometric model (include all the coincident factors) and determine all its characteristics the (parameters of the model, the mean square deviation of the parameters of the model, the dispersion and the mean square deviation of errors of the model, the coefficients of multiple correlation and determination) with the help of the Data Analysis add-in in Microsoft Excel.
2. Check the statistical significance of the model parameters. Check the model's adequacy with the help of the Fisher's criterion.
3. Adduce tables with the theoretical values of the dependent indicator and the values of the model's errors. Build a graph of the linear function. Build a histogram and a graph of the distribution of errors. Adduce grouping of data depending on the values of errors, give economical interpretation.
4. Find the forecasted value of the dependent variable Y_{pr} and the confidential intervals if there is available data about the future values of independent indicators $(X_{1pr}, X_{2pr}, X_{3pr})$.
5. Adduce a matrix of pair correlations for factorial features. Check the model for presence of multicollinearity (tight linear connection) between the factorial variables with the help of the Farrar – Glauber algorithm.
6. Exclude from the model the factors which have the least influence on the dependent variable or are interconnected with each other (use the results of the Student's criterion, the Farrar – Glauber algorithm and the coefficients of pair correlations). Determine all characteristics of a new regression; draw conclusions about its adequacy.

Guidelines

1. Launching Microsoft Excel and preparation of data.

Let's consider that there is available information about the values of the dependent variable *y* (GNP, mln UAH) and the independent variables: x_1 (labor resources expenditures, mln UAH), x_2 (basic funds expenditures, mln UAH), x_3 (flow-out of capital to the off-shore zone, mln UAH). After launching Microsoft Excel enter the initial data as shown in Fig. 3.1.

	A	B	C	D	E
1	№	y	x_1	x_2	x_3
2	1	51,2	9,8	6,8	0,65
3	2	54,7	10,1	10,5	0,82
4	3	65,6	10,9	12,2	1,04
5	4	53,3	10,4	12,8	1,53
6	5	72,7	12,1	13,4	1,94
7	6	75,7	12,8	13,8	1,75
8	7	85,6	13,2	14,4	1,99
9	8	91,5	13,6	14,6	0,92
10	9	96,7	12,3	16,5	0,95
11	10	102,6	13,5	18,2	0,9

Fig. 3.1. **The initial data**

2. Building a linear multifactorial econometric model and determining all its characteristics with the help of the Data Analysis add-in in Microsoft Excel.

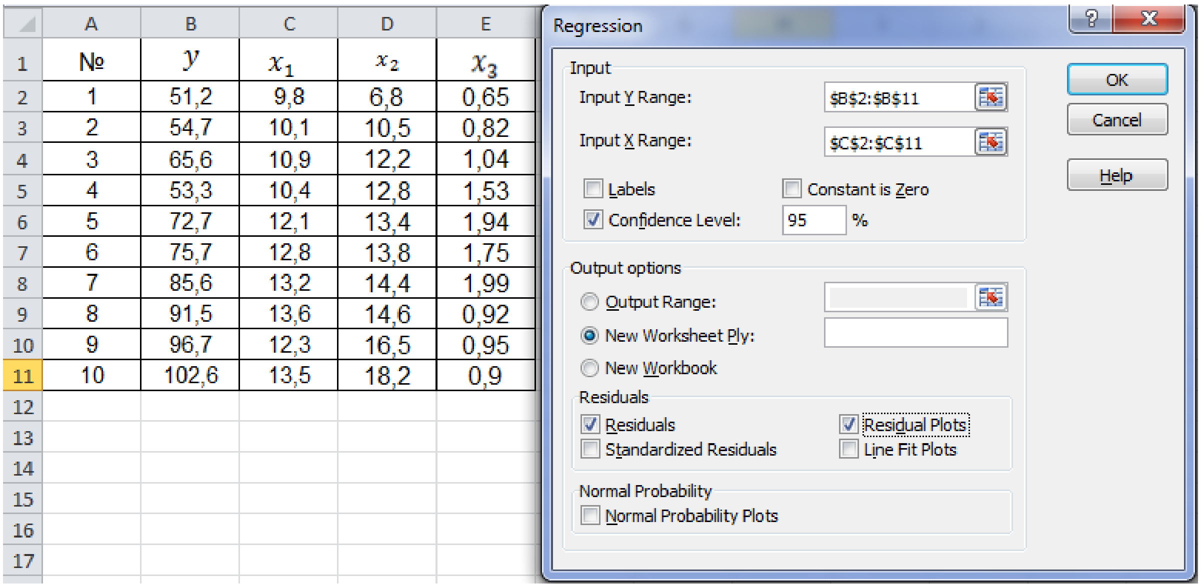
The parameters of linear multifactorial models and all their characteristics can be determined with the help of *the* Data Analysis/Regression add-in, the same way as for pair regression (see the guidelines for laboratory work 2), but choosing three adjoining columns x_1, x_2, x_3 as the output massive *X*, as shown in Fig. 3.2.

The results of building the multifactorial model are presented in Fig. 3.3.

The received model looks like:

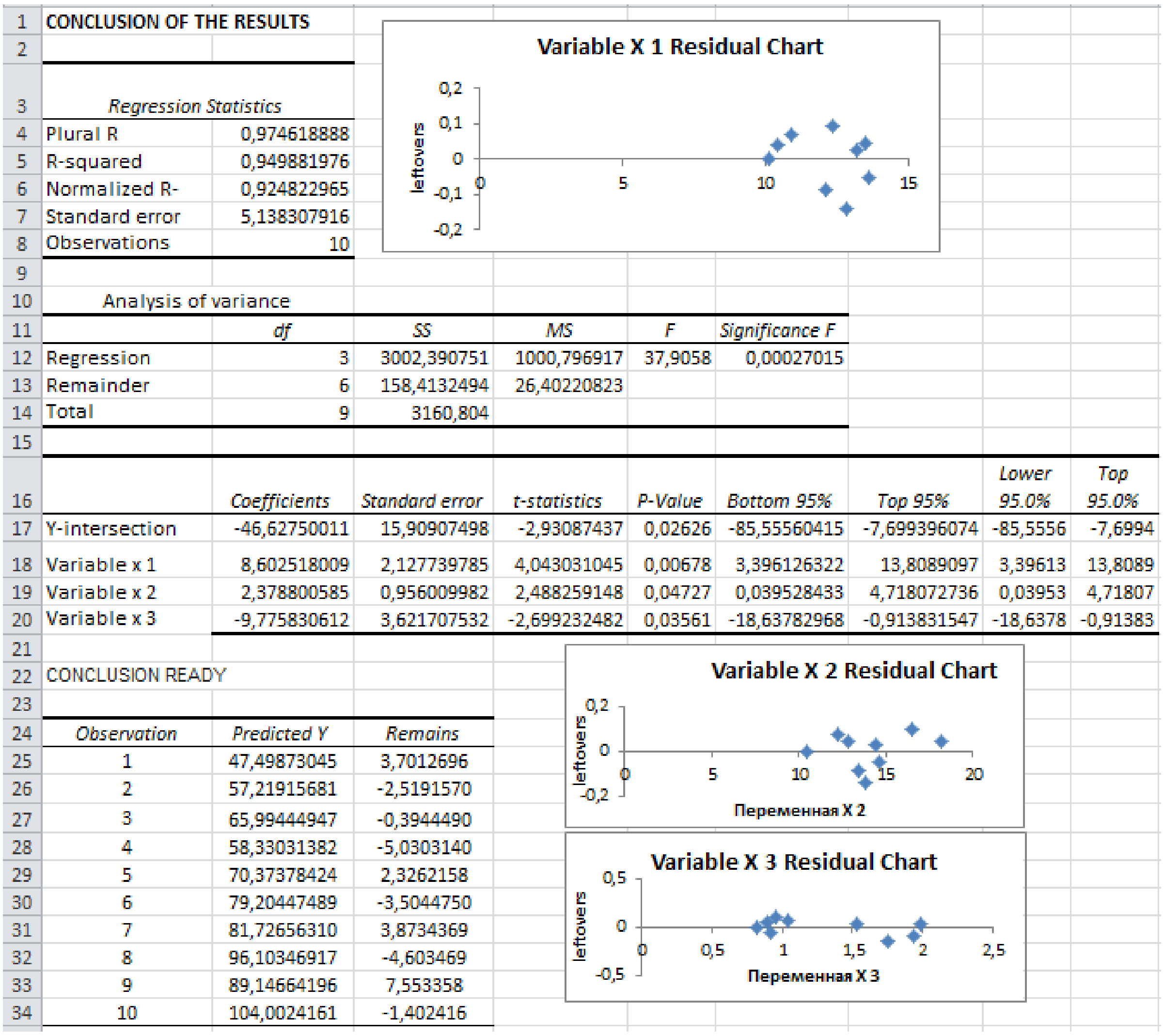
$$\hat{y} = -46.6275 + 8.6025 \cdot x_1 + 2.3788 \cdot x_2 - 9.7758 \cdot x_3 \qquad (18)$$

As can be seen in Fig. 3.3, the values of the coefficients of the model's adequacy (a more detailed description of which is given in the guidelines to practical activity 2) are quite high. The statistical significance of the model's parameters by the Student's criterion is: all parameters are significant ($t_{tabl}(0.05; 6) = 2.447$).



№	y	x_1	x_2	x_3
1	51,2	9,8	6,8	0,65
2	54,7	10,1	10,5	0,82
3	65,6	10,9	12,2	1,04
4	53,3	10,4	12,8	1,53
5	72,7	12,1	13,4	1,94
6	75,7	12,8	13,8	1,75
7	85,6	13,2	14,4	1,99
8	91,5	13,6	14,6	0,92
9	96,7	12,3	16,5	0,95
10	102,6	13,5	18,2	0,9

Fig. 3.2. **The dialog box of the Data Analysis/Regression add-in**



CONCLUSION OF THE RESULTS

Regression Statistics	
Plural R	0,974618888
R-squared	0,949881976
Normalized R-	0,924822965
Standard error	5,138307916
Observations	10

Analysis of variance

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	3002,390751	1000,796917	37,9058	0,00027015
Remainder	6	158,4132494	26,40220823		
Total	9	3160,804			

	Coefficients	*Standard error*	*t-statistics*	*P-Value*	*Bottom 95%*	*Top 95%*	*Lower 95.0%*	*Top 95.0%*
Y-intersection	-46,62750011	15,90907498	-2,93087437	0,02626	-85,55560415	-7,699396074	-85,5556	-7,6994
Variable x 1	8,602518009	2,127739785	4,043031045	0,00678	3,396126322	13,8089097	3,39613	13,8089
Variable x 2	2,378800585	0,956009982	2,488259148	0,04727	0,039528433	4,718072736	0,03953	4,71807
Variable x 3	-9,775830612	3,621707532	-2,699232482	0,03561	-18,63782968	-0,913831547	-18,6378	-0,91383

CONCLUSION READY

Observation	*Predicted Y*	*Remains*
1	47,49873045	3,7012696
2	57,21915681	-2,5191570
3	65,99444947	-0,3944490
4	58,33031382	-5,0303140
5	70,37378424	2,3262158
6	79,20447489	-3,5044750
7	81,72656310	3,8734369
8	96,10346917	-4,603469
9	89,14664196	7,553358
10	104,0024161	-1,402416

Fig. 3.3. **The results of building the multifactorial model with the help of the Data Analysis/Regression add-in**

3. Building a graph of the linear function.

To build a graph, it is necessary to copy the factual values of *Y* from the main window, and the theoretical values of $\hat{Y}$ (the column *Predicted Y*) from the window with the results of building a regression. Then, in the tab *Insert,* choose *Graph / Graph with markers* and enter the input data for building a graph. The result of building such a graph is presented in Fig. 3.4 (for the range of the factual values of *Y* the type of diagram was changed into Dotted).

№	y	Predict Y
1	51,2	47,49873
2	54,7	57,21916
3	65,6	65,99445
4	53,3	58,33031
5	72,7	70,37378
6	75,7	79,20447
7	85,6	81,72656
8	91,5	96,10347
9	96,7	89,14664
10	102,6	104,0024

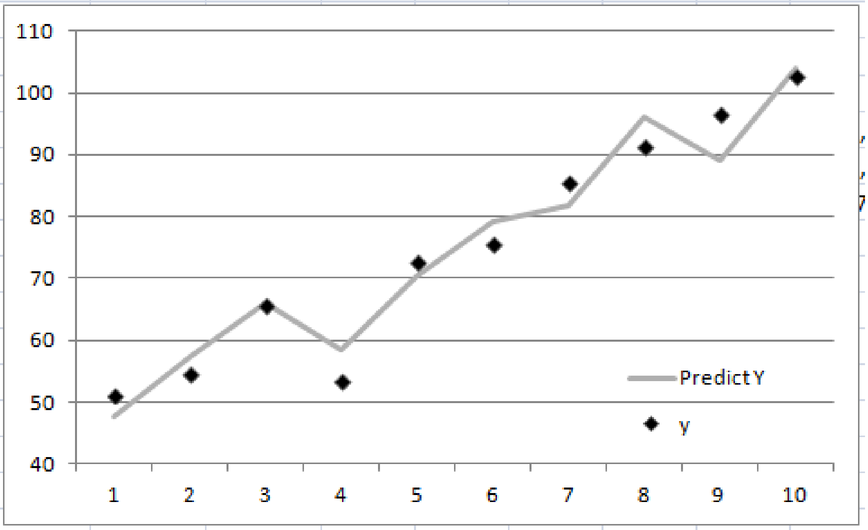


Fig. 3.4. **The results of building a multifactorial model with the help of the Data Analysis/Regression add-in**

So, the factual and theoretical values of the dependent variable on the graph (Fig. 3.4) are quite close, which means that the built model is of appropriate quality, but additional analytical procedures are necessary.

4. The graphs of the distribution of the model's residuals (errors).

Grouping the errors and building their histogram and the graph of distribution can be realized in the same way as for a simple linear model (see the guidelines to practical activity 2). The results of building are presented in Fig. 3.5.

It is seen from the graph of distribution (Fig. 3.5) that the model's errors significantly differ from zero and have quite a wide scope of values, which shows a rather low quality of the built model. Visual analysis of the histogram of the distribution of errors doesn't confirm the normal law of distribution.

Observation	Predicted Y	Remains
1	47,49873045	3,7012696
2	57,21915681	-2,5191570
3	65,99444947	-0,3944490
4	58,33031382	-5,0303140
5	70,37378424	2,3262158
6	79,20447489	-3,5044750
7	81,72656310	3,8734369
8	96,10346917	-4,603469
9	89,14664196	7,553358
10	104,0024161	-1,402416

max	7,553358
min	-5,030313821
swing	12,58367187
number of intervals	4
step	3,145917966

number of intervals	Lower boundary	Upper boundary	Frequency f
1	-5,030313821	-1,884395855	4
2	-1,884395855	1,261522112	2
3	1,261522112	4,407440078	3
4	4,407440078	7,553359044	1

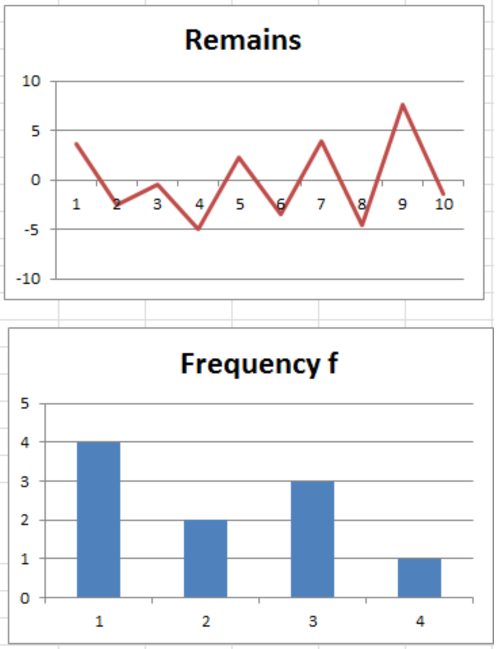


Fig. 3.5. **The histogram and the graph of the distribution of the model's errors**

5. The forecasted value of the dependent variable and the confidential intervals.

Calculation of the forecasted value of the dependent variable can be realized in the same way as in the simple linear model with the difference in the number of independent variables in the model and calculation Δy_{pr}.

$$y_{pr} - \Delta y_{pr} \le \tilde{y}_{pr} \le y_{pr} + \Delta y_{pr}, \qquad (19)$$

where $\Delta y_{pr} = t_p \cdot \sigma_e \sqrt{X_{pr}^T B^{-1} X_{pr}}$;

$$X_{pr}^T = (1, x_{1pr}, x_{2pr}, x_{3pr}).$$

To get the matrix $B^{-1} = \left(X^T X\right)^{-1}$ it is necessary first to get the matrices X and X^T, then to multiply them (X^TX) and find the inverse matrix $(X^TX)^{-1}$. All the formulas for calculations are given in Fig. 3.6.

The results of the calculations are presented in Fig. 3.7.

Let's consider that it is necessary to determine the forecasted value of GNP if the forecasted labor resources expenditures are 14 mln UAH, the basic funds expenditures make 17 mln UAH, and the out-flow of capital to off-shore zones amounts to 1.5 mln UAH. Then the vector of the forecasted independent variables will look like: X_{pr}^T (1, 14, 17, 1.5).

G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	1	=C2	=D2	=E2											
	1	=C3	=D3	=E3		=TRANSPOSE(H2:K11)	=TRANSPOSE(H2:K11)	=TRANS	=TRANS	=TRANS	=TRAN	=TRAN	=TRANSP	=TRANS	=TRANS
	1	=C4	=D4	=E4		=TRANSPOSE(H2:K11)	=TRANSPOSE(H2:K11)	=TRANS	=TRANS	=TRANS	=TRAN	=TRAN	=TRANSP	=TRANS	=TRANS
$X=$	1	=C5	=D5	=E5	$X^T=$	=TRANSPOSE(H2:K11)	=TRANSPOSE(H2:K11)	=TRANS	=TRANS	=TRANS	=TRAN	=TRAN	=TRANSP	=TRANS	=TRANS
	1	=C6	=D6	=E6		=TRANSPOSE(H2:K11)	=TRANSPOSE(H2:K11)	=TRANS	=TRANS	=TRANS	=TRAN	=TRAN	=TRANSP	=TRANS	=TRANS
	1	=C7	=D7	=E7											
	1	=C8	=D8	=E8			=MMULT(M3:V6;H2:K11)	=MMUL	=MMUL	=MMUL					
	1	=C9	=D9	=E9			=MMULT(M3:V6;H2:K11)	=MMUL	=MMUL	=MMUL					
	1	=C10	=D10	=E10		$X^TX=$	=MMULT(M3:V6;H2:K11)	=MMUL	=MMUL	=MMUL					
	1	=C11	=D11	=E11			=MMULT(M3:V6;H2:K11)	=MMUL	=MMUL	=MMUL					
							=MINVERSE(N8:Q11)	=MINVE	=MINVE	=MINVE					
						$B^{-1}=(X^TX)^{-1}=$	=MINVERSE(N8:Q11)	=MINVE	=MINVE	=MINVE					
							=MINVERSE(N8:Q11)	=MINVE	=MINVE	=MINVE					
							=MINVERSE(N8:Q11)	=MINVE	=MINVE	=MINVE					

Fig. 3.6. **The formulas for calculations**

	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1																
2		1	9,8	6,8	0,65											
3		1	10,1	10,5	0,82		1	1	1	1	1	1	1	1	1	1
4		1	10,9	12,2	1,04		9,8	10,1	10,9	10,4	12,1	12,8	13,2	13,6	12,3	13,5
5		1	10,4	12,8	1,53	$X^T=$	6,8	10,5	12,2	12,8	13,4	13,8	14,4	14,6	16,5	18,2
6	$X=$	1	12,1	13,4	1,94		0,65	0,82	1,04	1,53	1,94	1,75	1,99	0,92	0,95	0,9
7		1	12,8	13,8	1,75											
8		1	13,2	14,4	1,99			10	118,7	133,2	12,49					
9		1	13,6	14,6	0,92		$X^TX=$	118,7	1428	1614,9	150,39					
10		1	12,3	16,5	0,95			133,2	1614,9	1863,2	169,59					
11		1	13,5	18,2	0,9			12,49	150,39	169,59	17,863					
12																
13								9,5863	-1,113	0,2686	0,114					
14							$B^{-1}=(X^TX)^{-1}=$	-1,113	0,1715	-0,062	-0,073					
15								0,2686	-0,062	0,0346	0,0096					
16								0,114	-0,073	0,0096	0,4968					

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

The point estimation of the forecast looks like:

$$\hat{y} = -46.6275 + 8.6025 \cdot 14 + 2.3788 \cdot 17 - 9.7758 \cdot 1.5 \approx \\ \approx 99.584\ (\text{mln UAH}). \tag{20}$$

The formulas for calculating the confidential interval of the forecast are given in Fig. 3.8.

The results of the calculations are presented in Fig. 3.9.

The interval for the forecast looks like:

$$99.5836 - 7.31677 \le \tilde{y}_{pr} \le 99.5836 - 7.31677 \\ 92.97 \le \tilde{y}_{pr} \le 106.9. \tag{21}$$

I	J	K	L	M	N	O
		$t_p =$	=TINV(0,05;6)			
		$\sigma_e =$	='Total PA'!B7			
		$X_{pr}^T =$	1	14	17	1,5
		$X_{pr}^T B^{-1} =$	=MMULT(L21:O21;N13:Q16)	=MMUL	=MMULT	=MM
		$X_{pr}^T B^{-1} X_{pr} =$	=MMULT(L23:O23;O25:O28)			1
					$X_{pr} =$	14
		$\Delta Y_{pr} =$	=L18*L19*SQRT(L25)			17
						1,5

Fig. 3.8. **The formulas for calculations**

I	J	K	L	M	N	O
		$t_p =$	2,44961			
		$\sigma_e =$	5,13831			
	$X_{pr}^T =$		1	14	17	1,5
	$X_{pr}^T B^{-1} =$		-1,2528	0,1171	-0,0033	0,052
	$X_{pr}^T B^{-1} X_{pr} =$		0,33866			1
					$X_{pr} =$	14
	$\Delta Y_{pr} =$		7,31677			17
						1,5
	92,267	≤	99,58361603	≤	106,9003814	

Fig. 3.9. **The results of the point and interval forecast based on the model**

So, in the forecast period, the GNP can be from 92.7 mln UAH to 106.9 mln UAH.

6. Assessment of the model's multicollinearity.

To assess the multicollinearity, first of all it is necessary to determine the matrix of pair correlations between all variables. In the Data Analysis add-in, choose the option *Correlation*, whose dialog box after entering the output range of variables looks like in Fig. 3.10.

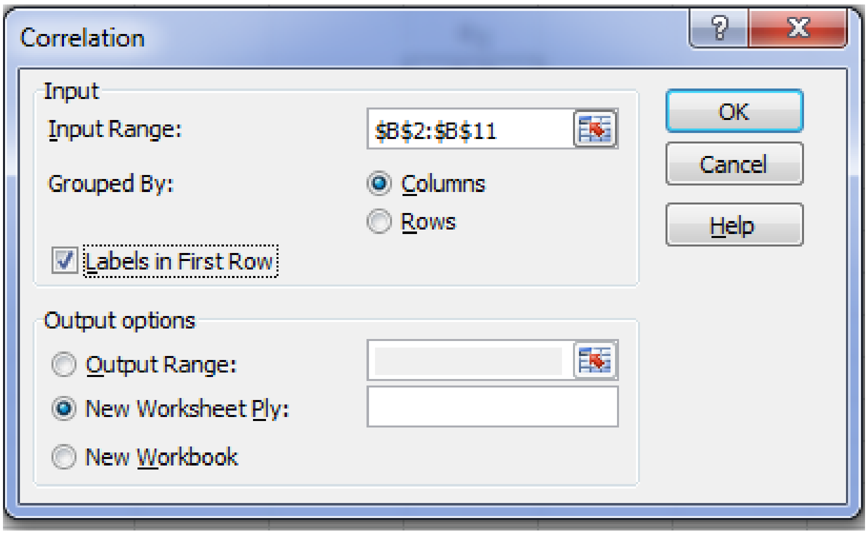


Fig. 3.10. **The dialog box of the tool *Correlation***

After confirmation of the output options with the *OK* button, the matrix of pair correlations will appear on a new page, shown in Fig. 3.11.

	A	B	C	D	E
1		y	x_1	x_2	x_3
2	y	1			
3	x_1	0,910209	1		
4	x_2	0,887585	0,820684	1	
5	x_3	0,046102	0,324931	0,227269	1

Fig. 3.11. **The matrix of pair correlations between the variables**

It is seen in Fig. 3.11, that Y is most influenced by X_1, and X_3 almost doesn't have any affect.

We also see, that there is a strong connection between the variables X_1 and X_2. This can tell us about the presence of multicollinearity in the model.

Let's check it with the help of the Farrar – Glauber algorithm. The results of the calculation are presented in Fig. 3.12.

The table value of the criterion χ^2 with the degrees of freedom $k = 0.5m(m-1) = 3$ and the level of significance $\alpha = 0.95$ is equal to $\chi^2_{tabl}(\alpha; 0.5m(m-1)) = 7.81$. Because $\chi^2 \geq \chi^2_{tabl}$, in the array of independent variables there is a general multicollinearity and the research needs to be continued.

	A	B	C	D	E
1		y	X_1	X_2	X_3
2	y	1			
3	X_1	0,910209100998135	1		
4	X_2	0,887584708057395	0,820683527259514	1	
5	X_3	0,046101659752674	0,32493130760063	0,22726942668450	1
6					
7		X_1	X_2	X_3	
8	X_1	1	0,820683	0,3249313	
9	X_2	0,8206	1	0,2272694	
10	X_3	0,32493	0,227269	1	
11					
12	det(r_{xx})=	=MDETERM(B8:D10)			
13	LN(det(r_{xx}))=	=LN(B12)			
14	χ^2=	=-(10-1-1/6*(2*3+5))*B13			
15	NDF=	=0,5*3*(3-1)			
16	χ^2 tabl	=CHIINV(0,05;B15)			

Fig. 3.12. **The results of the calculation**

Define the error matrix Z, inverse to the matrix $R = (r_{ij})$ using the function MINVERSE(array). The formula for calculation is shown in Fig. 3.13.

	A	B	C	D	E	F	G	H
7		X_1	X_2	X_3				
8	X_1	1	0,820683	0,3249313		=MINVERSE(B8:D10)	=MINVERS	=MINVER
9	X_2	0,8206	1	0,2272694	Z=R^{-1}=	=MINVERSE(B8:D10)	=MINVERS	=MINVER
10	X_3	0,32493	0,227269	1		=MINVERSE(B8:D10)	=MINVERS	=MINVER

Fig. 3.13. **The formula for calculating the matrix *Z***

We get the following result:

$$Z = R^{-1} = \begin{pmatrix} 3.2650 & -2.5712 & -0.4765 \\ -2.5712 & 3.0794 & 0.1356 \\ -0.4765 & 0.1356 & 1.1240 \end{pmatrix}. \qquad (22)$$

Next, we should calculate the multiple correlation coefficients R_i, which characterize the closeness of the connection of each variable with other variables:

$$R_1 = \sqrt{1 - \frac{1}{z_{11}}} = \sqrt{1 - \frac{1}{3.265}} \approx 0.833; \tag{23}$$

$$R_2 = \sqrt{1 - \frac{1}{z_{22}}} = \sqrt{1 - \frac{1}{3.0794}} \approx 0.822; \tag{24}$$

$$R_3 = \sqrt{1 - \frac{1}{z_{33}}} = \sqrt{1 - \frac{1}{1.124}} \approx 0.332. \tag{25}$$

As we see, the first and second regressors have a close connection with other regressors.

We should also check the statistical significance of the connection of each variable with other variables based on the *F*-criterion:

$$F_1 = \frac{(z_{11} - 1)(n - m)}{m - 1} = \frac{(3.265 - 1)(10 - 3)}{3 - 1} \approx 7.93; \tag{26}$$

$$F_2 = \frac{(z_{22} - 1)(n - m)}{m - 1} = \frac{(3.0794 - 1) \cdot 7}{2} \approx 7.28; \tag{27}$$

$$F_3 = \frac{(z_{33} - 1)(n - m)}{m - 1} = \frac{(1.1240 - 1) \cdot 7}{2} \approx 0.43. \tag{28}$$

The table value of *F*-statistics with the level of significance $\alpha = 0.95$ and degrees of freedom $k_1 = (m - 1) = 3 - 1 = 2$ and $k_2 = (n - m) = 10 - 3 = 7$ is equal to $F_{tab}(0.95; 2; 7) = 4.74$.

Since the calculated values $F_1 > F_{tab}$ and $F_2 > F_{tab}$, the regressors X_1 and X_2 respectively are multicollinear with others.

Next, we find the partial correlation coefficients:

$$r_{12}^p = \frac{-z_{12}}{\sqrt{z_{11} \cdot z_{22}}} = \frac{-(-2.5712)}{\sqrt{3.265 \cdot 3.0794}} \approx 0.811; \tag{29}$$

$$r_{13}^p = \frac{-z_{13}}{\sqrt{z_{11} \cdot z_{33}}} = \frac{-(-0.4765)}{\sqrt{3.265 \cdot 1.124}} \approx 0.249; \tag{30}$$

$$r_{23}^p = \frac{-z_{23}}{\sqrt{z_{22} \cdot z_{33}}} = \frac{-(0.1356)}{\sqrt{3.0794 \cdot 1.124}} \approx -0.07. \tag{31}$$

There is a close connection between the regressors X_1 and X_2.

It is interesting to compare the obtained values of the partial and pair correlation coefficients. Usually the former are much lower than the latter. But in this case, the partial coefficients that characterize the density of the relationship between the two variables, assuming that the other variables do not affect this relationship, are, in absolute terms, not significantly lower than those of the pairs, moreover for regressors X_2 and X_3 there was even a change in the direction of relation from the direct to the reverse. Thus, we can conclude that other variables significantly affect the relationship between the investigated indicators.

Next, we check the statistical significance of the relationship between each of the two variables based on the calculation of the *t*-criterion by the formula:

$$t_{12} = \left|r_{12}^{p}\right| \frac{\sqrt{n-m}}{\sqrt{1-\left(r_{12}^{p}\right)^2}} = |0.811| \cdot \frac{\sqrt{10-3}}{\sqrt{1-(0.811)^2}} \approx 3.67; \qquad (32)$$

$$t_{13} = \left|r_{13}^{p}\right| \frac{\sqrt{n-m}}{\sqrt{1-\left(r_{13}^{p}\right)^2}} = |0.249| \cdot \frac{\sqrt{7}}{\sqrt{1-(0.249)^2}} \approx 0.68; \qquad (33)$$

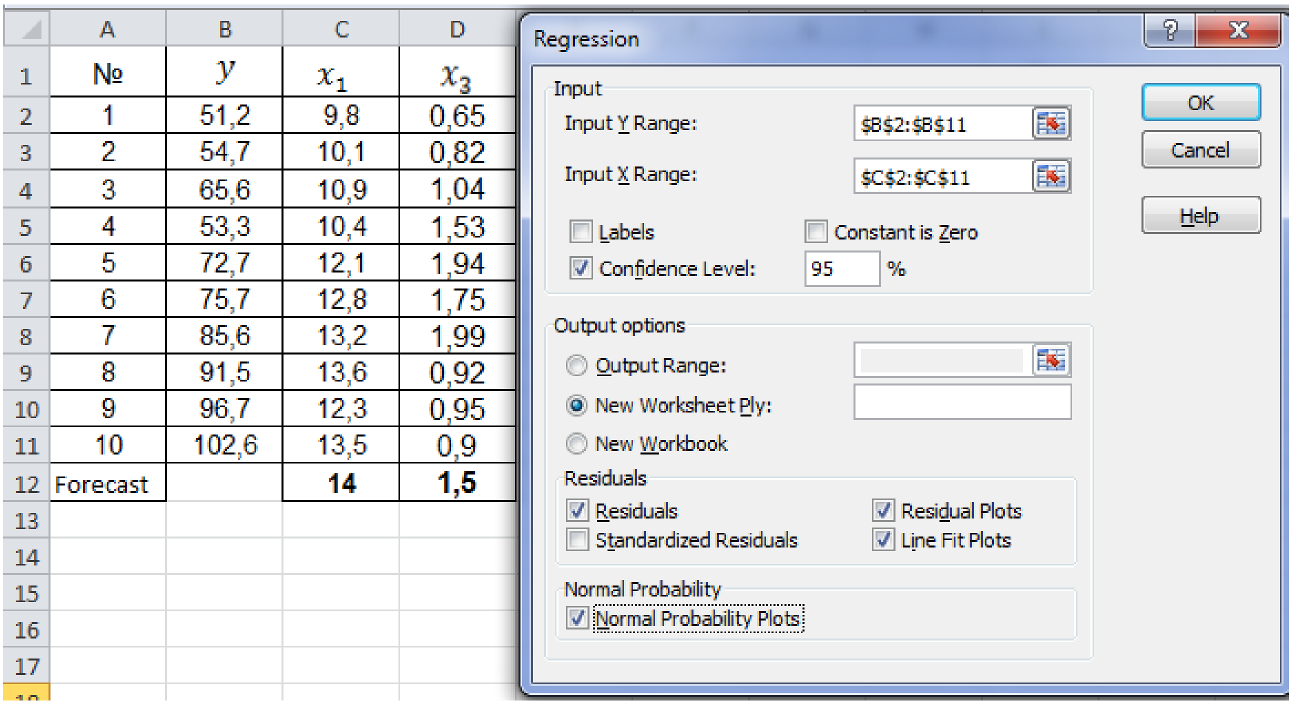
$$t_{23} = \left|r_{23}^{p}\right| \frac{\sqrt{n-m}}{\sqrt{1-\left(r_{23}^{p}\right)^2}} = |-0.07| \cdot \frac{\sqrt{7}}{\sqrt{1-(-0.07)^2}} \approx 0.193. \qquad (34)$$

The estimated values of the t-statistic are compared with the tabular ones with degrees of freedom $k = (n-m) = 10-3 = 7$ and the significance level $\alpha = 0.95$. Since $t_{12} \geq t_p(0.95; 7) = 2.36$, there is a statistically significant multicollinearity between the independent variables X_1 and X_2.

To exclude this negative phenomenon, it is necessary to eliminate the independent variable which has the least influence on the dependent variable *Y*. As we see, such variable is X_2.

To build a new regression, the raw data should be copied to a new worksheet. Remove the column X_2, run *Data analysis / Regression* (Fig. 3.14).

The results of the characteristics of the built new model without X_2 are shown in Fig. 3.15.



	A	B	C	D
1	№	y	x_1	x_3
2	1	51,2	9,8	0,65
3	2	54,7	10,1	0,82
4	3	65,6	10,9	1,04
5	4	53,3	10,4	1,53
6	5	72,7	12,1	1,94
7	6	75,7	12,8	1,75
8	7	85,6	13,2	1,99
9	8	91,5	13,6	0,92
10	9	96,7	12,3	0,95
11	10	102,6	13,5	0,9
12	Forecast		**14**	**1,5**
13				
14				
15				
16				
17				

Fig. 3.14. **The initial data and the dialog tool *Regression* for the new two-factor regression**

	A	B	C	D	E	F	G	H	I
1	CONCLUSION OF THE RESULTS								
3	*Regression Statistics*								
4	Plural R	0,947716							
5	R-squared	0,898165							
6	Normalized R-squared	0,869069							
7	Standard error	6,781074							
8	Observations	10							
10	Analysis of variance								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	2	2838,923	1419,462	30,869292	0,000337009			
13	Remainder	7	321,8808	45,98297					
14	Total	9	3160,804						
16		*Coefficients*	*Standard error*	*t-statistics*	*P-Value*	*Bottom 95%*	*Top 95%*	*Lower 95.0%*	*Top 95.0%*
17	Y-intersection	-65,082	18,57428	-3,50388	0,009941	-109,0032135	-21,1608	-109,003	-21,1608
18	Variable x 1	12,89576	1,643172	7,848088	0,000103	9,010272964	16,78124	9,010273	16,78124
19	Variable x 2	-10,4328	4,766883	-2,18861	0,0648119	-21,70472019	0,839054	-21,7047	0,839054

Fig. 3.15. **The characteristics of the new model after elimination of multicollinearity**

Fig. 3.15 shows that despite the insignificant decrease of coefficients of correlation and determination, in general, the quality of the model has become better after eliminating the variable X_2, in particular: all parameters of the model are statistically significant, the range of the interval assessments of the parameters and residuals of the model has decreased. So, for forecasting the dependent variable Y it is better to use the model:

$$\hat{Y} = -65.082 + 12.8958 \cdot X_1 - 10.4328 \cdot X_3. \qquad (35)$$

Content module 2
Applied econometrics

Practical activity 4. Testing the residuals for autocorrelation and heteroscedasticity

The goal is to consolidate the theoretical material and to acquire practical skills in testing an econometric model for the presence of autocorrelation and heteroscedasticity of the residuals.

In Table 4.1, the data describing the activities of 22 commercial banks, is presented.

Table 4.1

The initial data

Bank No.	The number of loan agreements (x_1)	Authorized capital (x_2)	Commercial bank revenue (y)	Bank No.	The number of loan agreements (x_1)	Authorized capital (x_2)	Commercial bank revenue (y)
1	24.0	468	61.6	12	52.8	552	109.7
2	26.4	456	68.6	13	54.0	576	102.1
3	27.6	456	64.7	14	55.2	624	117.8
4	30.0	492	75.8	15	58.8	612	109.9
5	32.4	504	73.3	16	62.4	540	121.3
6	33.6	504	81.4	17	68.4	552	116.3
7	36.0	660	88.2	18	70.8	540	132.2
8	38.4	624	96.0	19	76.8	600	128.9
9	40.8	492	81.7	20	79.2	660	151.6
10	42.0	540	94.8	21	81.6	708	141.7
11	44.4	564	90.7	22	84.0	960	178.8

The task is as follows:

1. Construct a linear multifactorial model, determine all its characteristics (the mean square deviations of the model parameter estimates, the Student criterion, the multiple correlation coefficient, the determination coefficient, the Fisher criterion). Draw conclusions about the statistical significance of the model.

2. Test the residuals of the model for the presence of autocorrelation using the Durbin – Watson criterion, the Von Neumann criterion and the cyclic coefficient of autocorrelation. Draw conclusions. Choose the most appropriate method for estimating the parameters of the econometric model.

3. Test the hypothesis for the presence of heteroscedasticity with the Park, Glaser, White tests. Draw conclusions. Choose the most appropriate method for estimating the parameters of the econometric model.

Guidelines

Autocorrelation of the residual is the presence of a relationship between the successive elements of a series of the model residuals. Autocorrelation of residuals is most often observed when the econometric model is based on a time series. If there is a correlation between the successive values of some explanatory variable, then the correlation of successive values of the residual will be observed. If the model with autocorrelation of residuals is evaluated by the use of LSMs, then the following negative consequences are possible: the estimation of the model parameters may be unmatched and substantiated but ineffective, i.e., the sample variance of the estimation of a_i may be unreasonably large; the statistical criterion *t*- and *F*-statistics can't be used to verify the model, since the calculation does not take into account the presence of correlations of the residuals; the ineffectiveness of estimations of the parameters of an econometric model usually results in ineffective predictions, that is, the predictive values will have a large sample dispersion.

1. Construction of the multiple regression.

With the help of the add-in Data Analysis, tabs Regression on the basis of data in Table 4.1, find the characteristics of the regression equation of the dependence of the banks' income (y) on the number of loan agreements (x_1) and the authorized capital (x_2), as shown in Fig. 4.1.

The results of the regression analysis are given in Fig. 4.2.

No.	X_1	X_2	Y
1	24	468	61.6
2	26.4	456	86.6
3	27.6	456	64.7
4	30	492	75.8
5	32.4	504	73.3
6	33.6	504	81.4
7	36	660	88.2
8	38.4	624	96
9	40.8	492	81.7
10	42	540	94.8
11	44.4	564	90.7
12	52.8	552	109.7
13	54	576	102.1
14	55.2	624	117.8

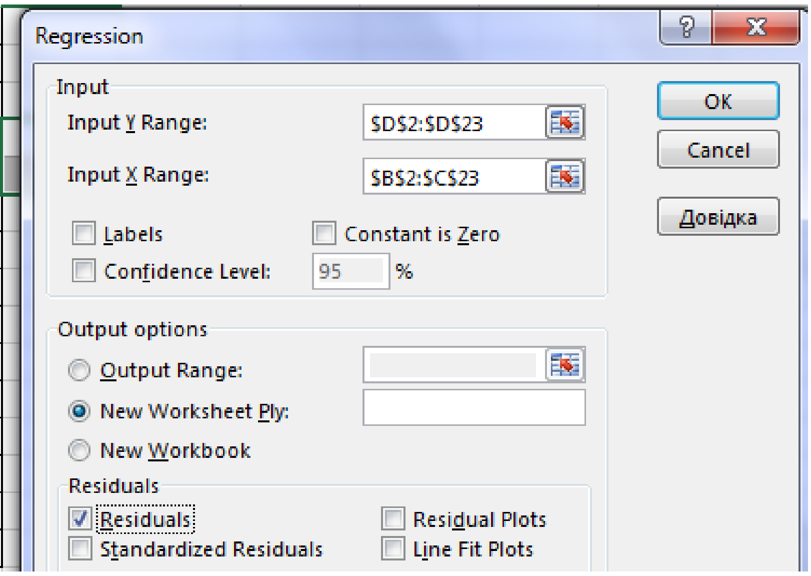


Fig. 4.1. **Calculation of the model parameters using the Data Analysis add-in, the Regression tab**

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	*Regression Statistics*						
4	Multiple R	0,98375444					
5	R Square	0,9677728					
6	Adjusted R Square	0,96438046					
7	Standard Error	5,72533501					
8	Observations	22					
9							
10	ANOVA						
11		*df*	*SS*	*MS*	*F*	*Significance F*	
12	Regression	2	18702,78	9351,392	285,282	6,73154E-15	
13	Residual	19	622,8098	32,77946			
14	Total	21	19325,59				
15							
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
17	Intercept	-3,42940094	6,852244	-0,50048	0,622486	-17,77131296	10,91251109
18	X Variable 1	1,16901406	0,088144	13,26252	4,7E-11	0,984526097	1,353502031
19	X Variable 2	0,08307464	0,01558	5,332182	3,81E-05	0,050465622	0,115683649

Fig. 4.2. **The results of the regression analysis**

Thus, after applying LSM to the initial data, the following regression equation is obtained:

$$y = -3.4294 + 1.17 \cdot x_1 + 0.083 \cdot x_2 + e; \quad R^2 = 0.968; \quad F = 285.282. \quad (36)$$

This equation is statistically significant with the probability 0.95 (F_{tab} = 3.52), the relationship between the indicators is very tight. The parameters of the factors are also significant (t_{fact} is equal to 13.26 and 5.33 when t_{tab} = 2.09). Thus, the considered characteristics indicate a high quality of the model. However, conclusions about the significance of the parameters will be reliable, and the model can be used for further analysis and forecast if the analysis of random residuals will not establish a violation of the LSM preconditions.

Here are the model residuals series (Fig. 4.3), obtained using the Data Analysis add-in, the Regression tab (Fig. 4.1).

23	RESIDUAL OUTPUT		
24			
25	*Observation*	*Predicted Y*	*Residuals*
26	1	63,5058659	-1,90587
27	2	65,3146041	3,285396
28	3	66,7174209	-2,01742
29	4	72,5137416	3,286258
30	5	76,3162709	-3,01627
31	6	77,7190878	3,680912
32	7	93,4843647	-5,28436
33	8	93,2993116	2,700688
34	9	85,1390935	-3,43909
35	10	90,5294928	4,270507
36	11	95,3289178	-4,62892
37	12	104,15174	5,54826
38	13	107,548348	-5,44835
39	14	112,938748	4,861252
40	15	116,150303	-6,2503
41	16	114,37738	6,92262
42	17	122,38836	-6,08836
43	18	124,197098	8,002902
44	19	136,19566	-7,29566
45	20	143,985772	7,614228
46	21	150,778989	-9,07899
47	22	174,51943	4,28057

Fig. 4.3. **The model residuals series**

2. Checking the residual for the presence of autocorrelation.

The most common methods of testing the residuals for the presence of autocorrelation are the Durbin – Watson criterion, the von Neumann criterion, the cyclic coefficient of autocorrelation. Let's consider the essence of these methods.

The Durbin – Watson criterion is based on testing the hypothesis about the existence of autocorrelation between the adjacent residual members of a series. The statistics that meet this criterion are usually labeled as *d* or *DW* and can be calculated by the formula:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}. \tag{37}$$

If auto-correlation is missing or small, then the value of *d* is approximately equal to 2, while with full autocorrelation the value of *d* is close to 0 or 4.

For *d*-statistics critical boundaries have been found that allow us to accept or reject the hypothesis about the presence of autocorrelation. The authors of this criterion estimated the lower (d_L) and upper (d_U) boundaries with 1; 2.5; 5 % levels of significance (Fig. 4.4):

1) if $0 < d \le d_L$, then the hypothesis about the presence of a positive autocorrelation is accepted;

2) if $d_L < d < d_U$, then there is no statistical reason to either accept or reject this hypothesis;

3) if $d_U \le d < 4 - d_U$, then the hypothesis about the absence of autocorrelation is accepted;

4) if $4 - d_U < d < 4 - d_L$, then there is no statistical reason to either accept or reject this hypothesis;

5) if $4 - d_L < d$ then there is a negative autocorrelation.

Positive autocorrelation	Uncertainty zone	Autocorrelation is absent	Uncertainty zone	Negative autocorrelation

0	d_L	d_U	2	$4 - d_U$	$4 - d_L$	4

Fig. 4.4. **Checking the hypothesis about the presence of autocorrelation of residuals according to the Durbin – Watson criterion**

In the case of the presence of a lagged dependent variable in the model, this criterion is not suitable. We can use the asymptotic Durbin *h*-test. Both of these tests are intended to test autocorrelation of random errors of the first order.

The Von Neumann criterion is determined by the formula:

$$Q = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2 \big/ n-1}{\sum_{t=1}^{n}(e_t)^2 \big/ n} = \frac{n}{n-1} \cdot \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2}. \tag{38}$$

The value of the Durbin – Watson statistics and the Von Neumann statistics are linked by the ratio: $Q = \frac{n}{n-1} \cdot d$.

If the value of Q is less (or more) than some critical value, then we can speak of positive (or negative) autocorrelation (Fig. 4.5).

Positive autocorrelation	Autocorrelation is absent	Negative autocorrelation

0 $\qquad Q_{cr}^{+}$ $\qquad Q_{cr}^{-}$ $\qquad \propto$

Fig. 4.5. **Verification of the hypothesis about the presence of autocorrelation of residuals by the Von Neumann criterion**

So, if $Q_{fact} < Q_{cr}^{+}$, then there is a positive autocorrelation, if $Q_{fact} > Q_{cr}^{-}$, then there is a negative autocorrelation, and if $Q_{cr}^{+} < Q_{fact} < Q_{cr}^{-}$, then the autocorrelation of the residuals is absent.

The cyclic coefficient of autocorrelation expresses the degree of interconnection of the "closed" series of residuals:

1st series – $e_1, e_2, e_3, \dots, e_{n-1}, e_n$;

2nd series – $e_2, e_3, e_4, \dots, e_n, e_1$.

It is calculated by the formula:

$$r^0 = \frac{\sum_{t=2}^{n} e_t e_{t-1} + e_n e_1 - \frac{1}{n}\left(\sum_{t=1}^{n} e_t\right)^2}{\sum_{t=1}^{n} e_t^2 - \frac{1}{n}\left(\sum_{t=1}^{n} e_t\right)^2}. \tag{39}$$

The calculated value of the cyclic coefficient of auto-correlation is compared with the table for the selected level of significance and the length of the row n (Fig. 4.6).

Negative autocorrelation	Autocorrelation is absent	Positive autocorrelation

-1 $\qquad -r_{cr}^{-}$ $\qquad 0$ $\qquad r_{cr}^{+}$ $\qquad +1$

Fig. 4.6. **Testing the hypothesis about the presence of autocorrelation of residuals by the cyclic autocorrelation coefficient**

If $r_{fact} > 0$ and $r_{fact} \geq r_{cr}^{+}$, then there is a positive autocorrelation. If $r_{fact} < 0$ and $r_{fact} \leq r_{cr}^{-}$, then there is a negative autocorrelation.

Assuming that:

$$\sum_{t=1}^{n} e_t \approx \sum_{t=2}^{n} e_{t-1} \approx 0, \tag{40}$$

the cyclic autocorrelation coefficient can be written as:

$$r^0 = \frac{n}{n-1} \cdot \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n} (e_t)^2}. \tag{41}$$

Let's calculate the value of the Durbin – Watson criterion, the Von Neumann criterion, the cyclic autocorrelation coefficient by the formulas (37) – (41). The results of the interim calculations are shown in Table 4.2.

Table 4.2

The interim calculations

No.	Y	$\hat{Y}$	Residuals, e_t	$(e_t - e_{t-1})^2$	e_t^2	$e_t \cdot e_{t-1}$
1	2	3	4	5	6	7
1	61.6	63.5059	-1.9059	---	3.6323	---
2	68.6	65.3146	3.2854	26.9492	10.7938	-6.2615
3	64.7	66.7174	-2.0174	28.1199	4.0700	-6.6280
4	75.8	72.5137	3.2863	28.1290	10.7995	-6.6298
5	73.3	76.3163	-3.0163	39.7219	9.0979	-9.9122
6	81.4	77.7191	3.6809	44.8523	13.5491	-11.1026
7	88.2	93.4844	-5.2844	80.3762	27.9245	-19.4513
8	96	93.2993	2.7007	63.7611	7.2937	-14.2714
9	81.7	85.1391	-3.4391	37.6969	11.8274	-9.2879
10	94.8	90.5295	4.2705	59.4379	18.2372	-14.6867
11	90.7	95.3289	-4.6289	79.1998	21.4269	-19.7678
12	109.7	104.152	5.5483	103.5749	30.7832	-25.6824
13	102.1	107.548	-5.4483	120.9254	29.6845	-30.2289
14	117.8	112.939	4.8613	106.2879	23.6318	-26.4858
15	109.9	116.15	-6.2503	123.4667	39.0663	-30.3843
16	121.3	114.377	6.9226	173.5259	47.9227	-43.2685

Table 4.2 (the end)

1	2	3	4	5	6	7
17	116.3	122.388	-6.0884	169.2856	37.0681	-42.1474
18	132.2	124.197	8.0029	198.5637	64.0464	-48.7245
19	128.9	136.196	-7.2957	234.0460	53.2267	-58.3865
20	151.6	143.986	7.6142	222.3048	57.9765	-55.5508
21	141.7	150.779	-9.0790	278.6635	82.4280	-69.1295
22	178.8	174.519	4.2806	178.4778	18.3233	-38.8632
∑	2287.1	2287.1	0.0000	2397.3662	622.8098	-586.8511

As a result of the calculation of the Durbin – Watson criterion, the following result is obtained:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{2\,397.3662}{622.8098} \approx 3.849. \qquad (42)$$

In the Durbin – Watson criterion there are upper and lower limits. For a model with two independent variables and 22 observations, the lower limit $d_L = 1.15$, the upper limit $d_U = 1.54$ with the level of significance 0.05 %.

Since the calculated value d gets into the interval $4 - d_L \le d < 4$ ($2.85 \le d < 4$), then we can conclude that there is a negative autocorrelation of the first order residuals in the model under study.

Let's calculate the Von Neumann statistics:

$$Q = \frac{n}{n-1} \cdot \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n}(e_t)^2} = \frac{n}{n-1} \cdot d = \frac{22}{21} \cdot 3.849 = 4.03. \qquad (43)$$

The critical values of this criterion for sample size $n = 22$ and the level of significance $\alpha = 0.05$ are: $Q_{cr}^{+} = 1.37$ and $Q_{cr}^{-} = 3.12$. Since the value $Q = 4.03$, $Q > 3.12$, there is a negative autocorrelation of the residuals.

Let's calculate the cyclic coefficient of autocorrelation:

$$r^0 = \frac{n}{n-1} \cdot \frac{\sum_{t=2}^{n} e_t e_{t-1}}{\sum_{t=1}^{n}(e_t)^2} = \frac{22 \cdot (-586.8511)}{21 \cdot 622.8098} = -0.987. \qquad (44)$$

The obtained value $r_{fact} < 0$. The critical value for negative autocorrelation for the sample size $n = 22$ and the significance level $\alpha = 0.05$ is: $r_{cr}^{-} = 0.399$. Since $|-0.987| > 0.399$, there is a negative autocorrelation of the residuals.

All three criteria bring us to the same conclusion.

In the case the hypothesis about the presence of autocorrelation is confirmed and the residuals can be represented as the first order autoregressive scheme, the Aitken method (GLSM) should be used to evaluate the model parameters. If the residuals can be represented as a higher order autoregressive scheme, the Cochrane – Orcutt and the Durbin methods are used.

3. Testing the heteroscedasticity of the residuals.

One of the important assumptions in constructing a regression model is that random errors in the model are uncorrelated with each other and have a constant variance. This requirement when using the usual least squares method for estimating the parameters of a general linear econometric model is called homoscedasticity. In practice, the requirement of a constant variance of random errors is often not fulfilled, and this phenomenon is called heteroscedasticity. When using a conventional OLS, the presence of heteroscedasticity of residuals will lead to a situation, when the model parameter estimates will be shifted, grounded, but ineffective.

Let's test the residual for heteroscedasticity.

The tests that can detect the presence of heteroscedastic residuals, include the Park, Glaser, White tests. These tests assume that the variability of random residuals is a function of dependence on some factor (or factors).

For the Park test, this dependence is as follows:

$$\ln e_i^2 = a + b \ln x_{ij} + v_i, \tag{45}$$

where x_{ij} is the value *i* for factor *j*;

v_i is the random residual.

The Glaser test should be used to find the parameters of a series of equations given by the function:

$$|e_i| = a + b x_{ij}^k + v_i, \tag{46}$$

where *k* is any number. For example, *k* = –1; –0.5; 0.5; 1, etc.

The White test is based on constructing a quadratic function that includes all factors, as well as their pairwise products. In particular, for a case with two factors, this function will have the form:

$$e_i^2 = a + b_{11}x_{1i} + b_{12}x_{1i}^2 + b_{21}x_{2i} + b_{22}x_{2i}^2 + c_{12}x_{1i}x_{2i} + v_i. \tag{47}$$

The residuals are considered heteroscedastic if the parameter *b* in the functions of the Park test (45) or the Glaser test (46) is statistically significant (for the Glaser test – at least at one *k* value). While conducting the White test, the heteroscedasticity of random residuals is given if the entire function (47) is significant by the Fisher's criterion F.

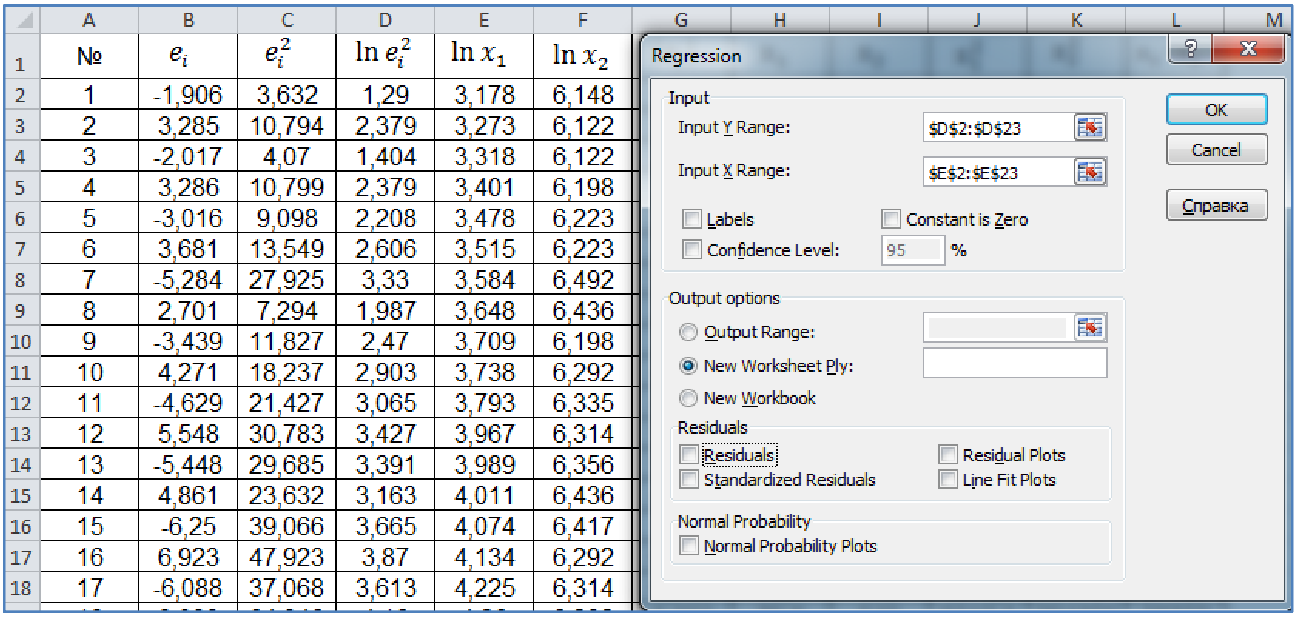
Let's calculate the values of dependent and independent variables of the functions (45) – (47). The results of the interim calculations are shown in Table 4.3.

Table 4.3

The interim calculations

No.	e_i	e_i^2	$\ln e_i^2$	$\ln x_1$	$\ln x_2$	$\lvert e_i \rvert$	x_1	x_2	x_1^2	x_2^2	$x_1 \cdot x_2$
1	-1.906	3.632	1.290	3.178	6.148	1.906	24	468	576	219024	11232
2	3.285	10.794	2.379	3.273	6.122	3.285	26.4	456	696.96	207936	12038
3	-2.017	4.070	1.404	3.318	6.122	2.017	27.6	456	761.76	207936	12586
4	3.286	10.799	2.379	3.401	6.198	3.286	30	492	900	242064	14760
5	-3.016	9.098	2.208	3.478	6.223	3.016	32.4	504	1049.8	254016	16330
6	3.681	13.549	2.606	3.515	6.223	3.681	33.6	504	1129	254016	16934
7	-5.284	27.925	3.330	3.584	6.492	5.284	36	660	1296	435600	23760
8	2.701	7.294	1.987	3.648	6.436	2.701	38.4	624	1474.6	389376	23962
9	-3.439	11.827	2.470	3.709	6.198	3.439	40.8	492	1664.6	242064	20074
10	4.271	18.237	2.903	3.738	6.292	4.271	42	540	1764	291600	22680
11	-4.629	21.427	3.065	3.793	6.335	4.629	44.4	564	1971.4	318096	25042
12	5.548	30.783	3.427	3.967	6.314	5.548	52.8	552	2787.8	304704	29146
13	-5.448	29.685	3.391	3.989	6.356	5.448	54	576	2916	331776	31104
14	4.861	23.632	3.163	4.011	6.436	4.861	55.2	624	3047	389376	34445
15	-6.250	39.066	3.665	4.074	6.417	6.250	58.8	612	3457.4	374544	35986
16	6.923	47.923	3.870	4.134	6.292	6.923	62.4	540	3893.8	291600	33696
17	-6.088	37.068	3.613	4.225	6.314	6.088	68.4	552	4678.6	304704	37757
18	8.003	64.046	4.160	4.260	6.292	8.003	70.8	540	5012.6	291600	38232
19	-7.296	53.227	3.975	4.341	6.397	7.296	76.8	600	5898.2	360000	46080
20	7.614	57.976	4.060	4.372	6.492	7.614	79.2	660	6272.6	435600	52272
21	-9.079	82.428	4.412	4.402	6.562	9.079	81.6	708	6658.6	501264	57773
22	4.281	18.323	2.908	4.431	6.867	4.281	84	960	7056	921600	80640

Let's test the residual for heteroscedasticity with the Park test (see formula (45)) for x_1. We find the parameters of the regression equation using the Data Analysis add-in, the Regression tab, as shown in Fig. 4.7.



	A	B	C	D	E	F
1	№	e_i	e_i^2	$\ln e_i^2$	$\ln x_1$	$\ln x_2$
2	1	-1,906	3,632	1,29	3,178	6,148
3	2	3,285	10,794	2,379	3,273	6,122
4	3	-2,017	4,07	1,404	3,318	6,122
5	4	3,286	10,799	2,379	3,401	6,198
6	5	-3,016	9,098	2,208	3,478	6,223
7	6	3,681	13,549	2,606	3,515	6,223
8	7	-5,284	27,925	3,33	3,584	6,492
9	8	2,701	7,294	1,987	3,648	6,436
10	9	-3,439	11,827	2,47	3,709	6,198
11	10	4,271	18,237	2,903	3,738	6,292
12	11	-4,629	21,427	3,065	3,793	6,335
13	12	5,548	30,783	3,427	3,967	6,314
14	13	-5,448	29,685	3,391	3,989	6,356
15	14	4,861	23,632	3,163	4,011	6,436
16	15	-6,25	39,066	3,665	4,074	6,417
17	16	6,923	47,923	3,87	4,134	6,292
18	17	-6,088	37,068	3,613	4,225	6,314

Fig. 4.7. **Calculation of regression parameters for the Park test**

The results of the regression analysis are shown in Fig. 4.8.

	A	B	C	D	E	F
1	SUMMARY OUTPUT					
2						
3	*Regression Statistics*					
4	Multiple R	0,862592897				
5	R Square	0,744066507				
6	Adjusted R Square	0,731269832				
7	Standard Error	0,449557335				
8	Observations	22				
9						
10	ANOVA					
11		*df*	*SS*	*MS*	*F*	*Significance F*
12	Regression	1	11,75127	11,75127	58,1453	2,42604E-07
13	Residual	20	4,042036	0,202102		
14	Total	21	15,79331			
15						
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
17	Intercept	-4,224658894	0,956238	-4,418	0,000265	-6,219335941
18	X Variable 1	1,881254295	0,246712	7,625307	2,43E-07	1,366622179

Fig. 4.8. **The results of the regression analysis**

Thus, the following regression equation is obtained:

$$\ln e^2 = -4.22 + 1.88 \ln x_1 + v; \quad t_b = 7.6. \tag{48}$$

Find the table value of Student's criterion, as shown in Fig. 4.9.

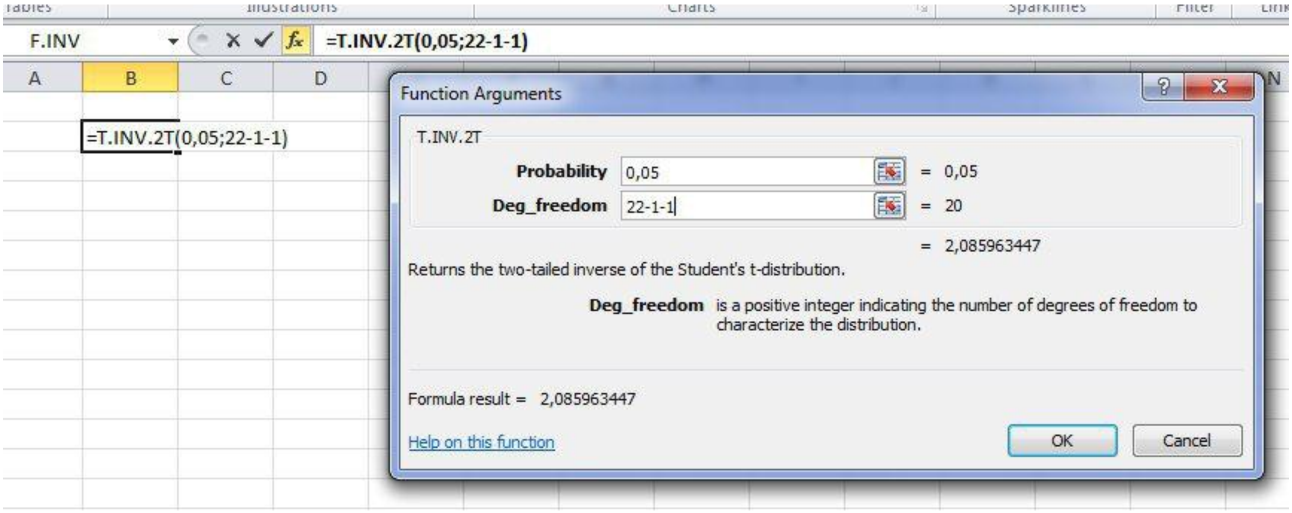


Fig. 4.9. **Determination of the critical value of Student's criterion**

Comparison of the calculated value of Student's criterion $t_b = 7.6$ and $t_{tab} = 2.09$ (Fig. 4.9) allows us to conclude that the parameter b is statistically significant, that is, the residuals of the model are heteroscedastic.

Similarly, we find the regression equation (45) for the Park test for the variable x_2, as well as the regression equation (46) by the Glaser test, the regression equation (47) by the White test. Below are the results of the regression analysis.

By the Park test:

$$\ln e^2 = -12.37 + 2.43 \ln x_2 + v; \quad t_b = 2.45. \tag{49}$$

According to the Glaser test, with $k = 1$:

$$|e| = 0.5894 + 0.0858 \cdot x_1 + \nu; \quad t_b = 6.91. \tag{50}$$

$$|e| = 1.122 + 0.0067 \cdot x_2 + \nu; \quad t_b = 1.77. \tag{51}$$

By the White test:

$$e^2 = -128.595 - 0.552 \cdot x_1 + 0.023 \cdot x_1^2 + 0.485 \cdot x_2 -$$
$$- 0.0003 \cdot x_2^2 - 0.0014 \cdot x_1 x_2 + \nu, \quad (52)$$
$$F = 23.665.$$

On the basis of the built-in function of Excel, we find the tabular value of Fisher's criterion, as shown in Fig. 4.10.

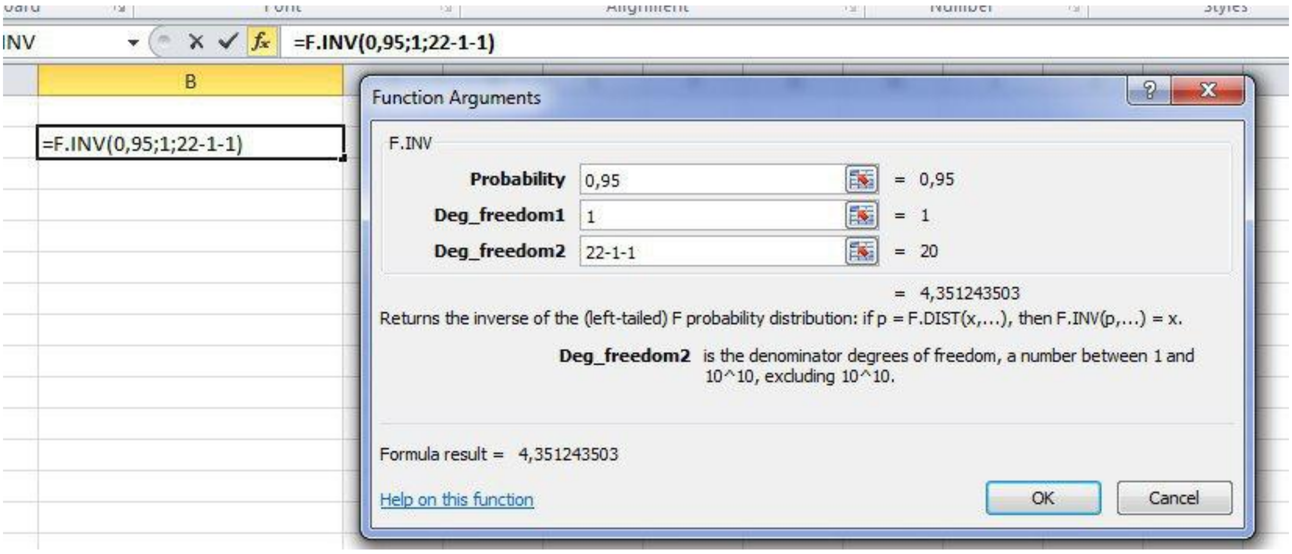


Fig. 4.10. **Determination of the critical value of Fisher's criterion**

The regression equation in the White test is significant (F_{tab} = 4.35). The parameters of the model in the Park test are significant (t_{tab} = 2.09), consequently, we can assert that the residuals are heteroscedastic. In the Glaser test, the parameter for the factor x_1 is significant, for the factor x_2 is insignificant, which allows us to conclude that the dispersion of residuals depends on the factor x_1.

In cases of violated requirements of the OLS concerning the nature of random residuals, namely, the constancy of the dispersion of random residuals, the non-correlation of the residuals among themselves, the generalized least squares method (GLS) is used.

The essence of this method is the elimination of the violation of the prerequisites of the OLS, adjusting the calculations of the parameters of the regression equation, taking into account the values of the covariance matrix of the residuals. Such a correction can be made using the formula:

$$\hat{a} = \left(X\ \Omega^{-1\prime} X\right)^{-1} X'\Omega^{-1} Y, \quad (53)$$

where Ω is the covariance matrix of residuals.

Practical activity 5. Evaluation and analysis of the main characteristics of the Cobb – Douglas production function

The goal of the study is to consolidate the theoretical material and obtain practical skills in assessing and analyzing the main properties and characteristics of the production function for the study of real economic processes.

The results of the population numbers (L), the volume of fixed assets (F), GDP (Y) are given in Table 5.1.

Table 5.1

The output data

No.	X_1, L (thousand people)	X_2, F (bln UAH)	Y, GDP (bln UAH)
1	1.3	2.2	2.15
2	2.5	2.6	4.41
3	2.7	3.3	5.54
4	2.9	3.6	6.69
5	3.5	4.6	7.48
6	4.5	5.2	9.56
7	6.1	5.8	10.62
8	7.2	6.2	11.94
9	8.6	7.5	12.02
10	8.4	10.8	18.51

The task is as follows:

1. Check if there is a non-linear relation between the volume of production and the value of production resources by constructing the Cobb – Douglas production function. Calculate the correlation index. Draw a conclusion concerning the adequacy of the non-linear econometric model.

2. Determine the characteristics of the production function (the average and marginal resource productivity, the elasticity of the product output based on the factors and the total elasticity, the capital-labour ratio).

3. Construct the isoquants of the production function, calculate the marginal norms of the substitution of resources at a given point on the isoquants. Draw conclusions.

Guidelines

The production function (PF) is a function in which the independent variable takes the value of the amount of the consumed or used resource (the factor of production), and the dependent variable is the value of the output volumes.

Production functions include modeling the dependencies that exist between the indicators of production activity, such as the volume of the output, the production costs, the capital expenditures, the capital productivity.

It is assumed that the output factors are the main productive assets X_1 and labour resources X_2.

The production function of Cobb – Douglas is an example of a concrete form of a two-factor function:

$$Y = a_0 \cdot X_1^{a_1} \cdot X_2^{a_2}, \tag{54}$$

where a_0, a_1 and a_2 are the parameters of the model.

1. **Constructing the Cobb – Douglas production function.** To linearize the linear dependence to evaluate the parameters of the PF of Cobb – Douglas the following formula is used:

$$\ln Y = \ln a_0 + a_1^* X_1 + a_2^* X_2. \tag{55}$$

Replace $\ln Y$ with Z, $\ln a_0$ with a_0^*, $\ln X_1$ with Z_1, $\ln X_2$ with Z_2, a_0 with, $e^{a_0^*}$, $Z = a_0^* + a_1^* Z_1 + a_2^* Z_2$.

The sequence of calculation of the model parameters $Z = a_0^* + a_1^* Z_1 + a_2^* Z_2$ in Excel is shown in Fig. 5.1.

When calculating the model parameters, the following built-in Excel functions are used: *MMULT(), MINVERSE()*. Based on the data shown in Fig. 5.1, the parameters of the model $Z = a_0^* + a_1 \cdot Z_1 + a_2 \cdot Z_2$ are equal to a_0^* = 0.42; a_1 = 0.52; a_2 = 0.57. That is, the modified model has the form: $Z = 0.42 + 0.52 \cdot Z_1 + 0.57 \cdot Z_2$. A similar result can be obtained using the Data Analysis add-in, the Regression tab, as shown in Fig. 5.2.

Step 1.

$z_1 = \ln(L)$	$z_2 = \ln(K)$	$z^* = \ln(Y)$
0.262	0.788	0.765
0.916	0.956	1.484
0.993	1.194	1.712
1.065	1.281	1.901
1.253	1.526	2.012
1.504	1.649	2.258
1.808	1.758	2.363
1.974	1.825	2.480
2.152	2.015	2.487
2.128	2.380	2.918

Step 2.

$Z =$

1	0.262	0.788
1	0.916	0.956
1	0.993	1.194
1	1.065	1.281
1	1.253	1.526
1	1.504	1.649
1	1.808	1.758
1	1.974	1.825
1	2.152	2.015
1	2.128	2.380

Step 3.

$Z^T Z =$

10	14.06	15.37
14.06	23.19	24.20
15.37	24.20	25.79

Step 4.

$(Z^T Z) =$

1.82	1.43	–2.43
1.43	3.25	–3.91
–2.43	-3.91	5.16

Step 5.

$\overline{z^* Z} =$

20.38
31.93
33.93

Step 6.

$\overline{a^*}$

0.42
0.52
0.57

$Z^T =$

1	1	1	1	1	1	1	1	1	1
0.2624	0.9163	0.9933	1.0647	1.2528	1.5041	1.8083	1.9741	2.1518	2.1282
0.7885	0.9555	1.1939	1.2809	1.5261	1.6487	1.7579	1.8245	2.0149	2.3795

Fig. 5.1.**The procedure for calculating the model parameters**

	z1=ln(L)	z2=ln(K)	z*=ln(Y)
1	0.262	0.788	0.765
2	0.916	0.956	1.484
3	0.993	1.194	1.712
4	1.065	1.281	1.901
5	1.253	1.526	2.012
6	1.504	1.649	2.258
7	1.808	1.758	2.363
8	1.974	1.825	2.480
9	2.152	2.015	2.487
10	2.128	2.380	2.918

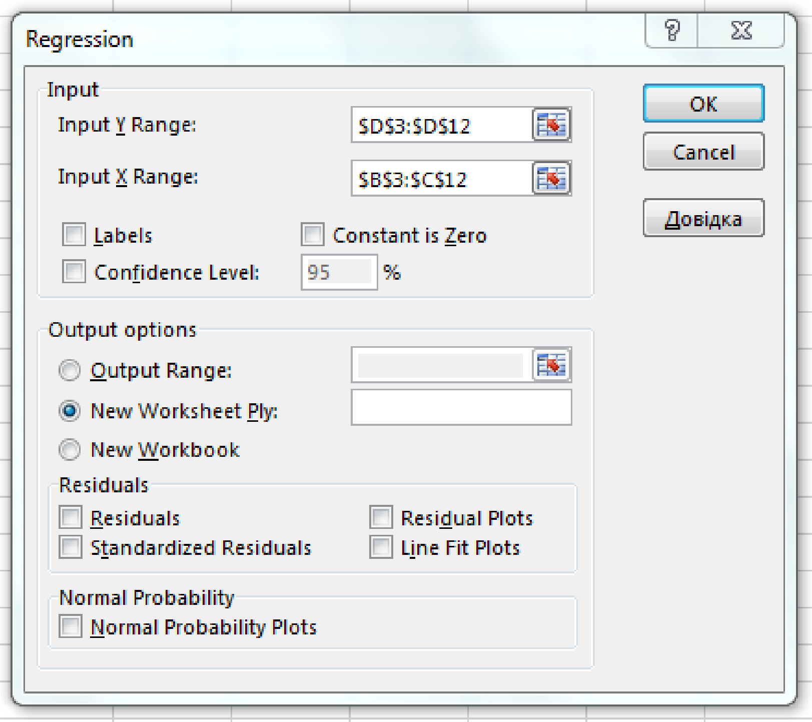


Fig. 5.2. **Calculation of the model parameters using the Data Analysis add-in, the Regression tab**

The results of the regression analysis are shown in Fig. 5.3.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,97510923
R Square	0,950838011
Adjusted R Square	0,936791729
Standard Error	0,154039203
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	3,212459467	1,60623	67,69321	2,63453E-05
Residual	7	0,166096533	0,023728		
Total	9	3,378556			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
Intercept	0,420838801	0,208080238	2,022483	0,082826	-0,071192776	0,912870377	-0,071192776	0,912870377
X Variable 1	0,522998516	0,278089876	1,880682	0,102064	-0,13457955	1,180576581	-0,13457955	1,180576581
X Variable 2	0,573826949	0,350038726	1,639324	0,145154	-0,253883112	1,401537009	-0,253883112	1,401537009

Fig. 5.3. **The results of the regression analysis**

The reverse transition to the nonlinear form of the model is performed $a_0^* = \ln a_0 \rightarrow a_0 = e^{a_0^*} = e^{0.42} = 1.52$ based on the built-in function *EXP().*

Thus, the constructed model of the production function of Cobb – Douglas has the form:

$$Y = 1.52 \cdot L^{0.52} K^{0.57}. \quad (56)$$

The quality of the resulting model is assessed based on the index of correlation. The results of the interim calculations are shown in Table 5.2.

Table 5.2

The interim calculations

No.	$\hat{y}$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y}_i)^2$
1	2.75	0.36	45.45
2	4.26	0.02	20.09
3	5.08	0.21	11.24
4	5.54	1.31	4.85
5	7.04	0.19	1.99
6	8.62	0.89	0.45
7	10.76	0.02	2.99
8	12.19	0.06	9.29
9	14.92	8.40	9.78
10	18.17	0.12	92.51
∑	-	11.59	198.63

The correlation index is:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{10} (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{10} (Y_i - \bar{Y})^2}} = \sqrt{1 - \frac{11.59}{198.63}} = 0.94. \quad (57)$$

The value of the correlation index which is equal to 0.94, allows us to draw a conclusion about the statistical significance of the model and the possibility of using it for further analysis.

2. Calculation of the characteristics of the production function at the point ($L_0 = 100, K_0 = 100$).

2.1. The value of the output for initial conditions with $L_0 = 100, K_0 = 100$ is calculated as:

$$Y = 1.52 \cdot 100^{0.52} \cdot 100^{0.57} = 230.06. \quad (58)$$

2.2. The average productivity of resources which shows the average number of products per unit of the spent labor is calculated according to the formula:

$$A_1 = \frac{y}{x_1} = a_0 \cdot x_1^{a_1 - 1} \cdot x_2^{a_2}; \qquad (59)$$

$$A_1 = 1.52 \cdot 100^{0.52-1} \cdot 100^{0.57} = 2.3. \qquad (60)$$

The average return on capital (the capital productivity) shows the volume of products per unit of the used production assets:

$$A_2 = \frac{y}{x_2} = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2 - 1}; \qquad (61)$$

$$A_2 = 1.52 \cdot 100^{0.52} \cdot 100^{0.57-1} = 2.3. \qquad (62)$$

Consider the geometric content of this characteristic. The average productivity of the resource is equal to the tangent of the angle of inclination of the chord, to the abscissa axis (Fig. 5.4), i.e. $A = \mathrm{tg}\alpha$, $\alpha = \mathrm{arctg}A$.

2.3. The marginal productivity of resources shows how many additional units of production an additional unit of the spent labor brings. It is calculated according to the formula:

$$M_1 = \frac{\partial y}{\partial x_1} = a_0 \cdot a_1 \cdot x_1^{a_1 - 1} \cdot x_2^{a_2}; \qquad (63)$$

$$M_1 = \frac{\partial y}{\partial x_1} = 1.52 \cdot 0.52 \cdot 100^{0.52-1} \cdot 100^{0.57} = 1.2. \qquad (64)$$

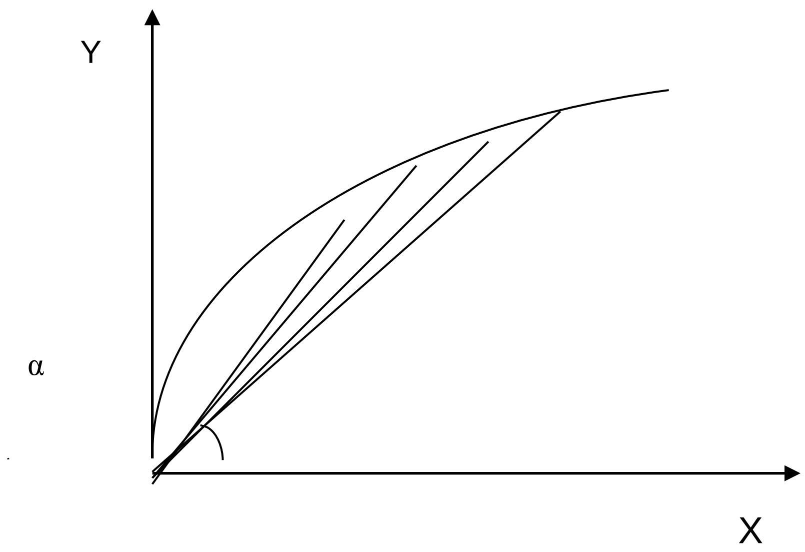


Fig. 5.4. **The graph of the resource average productivity**

The marginal return on capital (the capital outflow) shows how many additional units of production an additional unit of fixed assets brings.

$$M_2 = \frac{\partial y}{\partial x_2} = a_0 \cdot a_2 \cdot x_1^{a_1} \cdot x_2^{a_2 - 1}; \qquad (65)$$

$$M_2 = \frac{\partial y}{\partial x_1} = 1.52 \cdot 0.57 \cdot 100^{0.52} \cdot 100^{0.57-1} = 1.31. \qquad (66)$$

Consider the geometric content of this characteristic. The marginal efficiency (the productivity) of the resource is equal to the tangent of the angle of the tilt in the graph of the PF at the point x_0 to the axis of abscissa (Fig. 5.5), i.e. $M = \mathrm{tg}\varphi,\ \varphi = \mathrm{arctg}M$.

2.4. The elasticity of the production output as to the factors of production.

The elasticity of the production output in terms of labor costs shows how much the product output will increase with an increase in labor costs by 1 %. It can be defined as follows:

$$E_1 = \frac{\partial y}{\partial x_1} \cdot \frac{x_1}{y}; \quad E_1 = \frac{M_1}{A_1}; \quad E_1 = a_1; \qquad (67)$$

$$E_1 = \frac{1.2}{2.3} = 0.52; \quad a_1 = 0.52. \qquad (68)$$

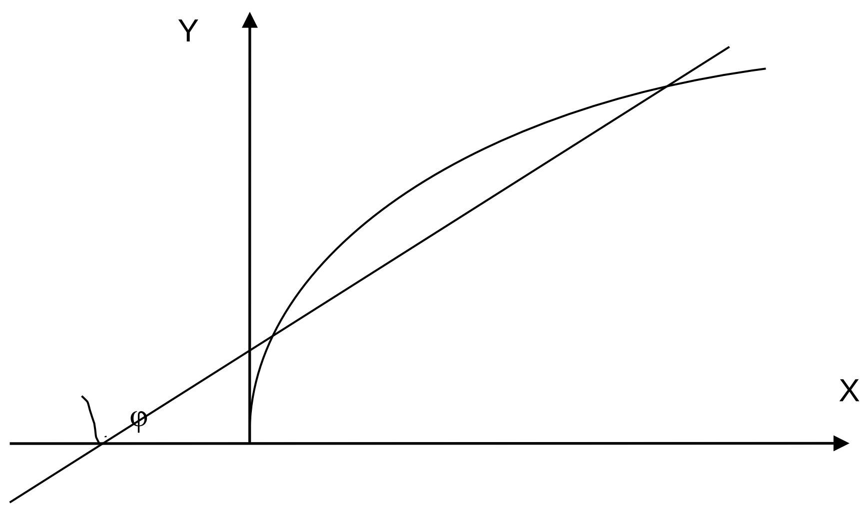


Fig. 5.5. **The graph of the marginal efficiency of the resource**

The elasticity of the production output in terms of the cost of production assets shows how much the percent of the production output will increase with an increase in fixed assets by 1 %.

$$E_2 = \frac{\partial y}{\partial x_2} \cdot \frac{x_2}{y}; \quad E_2 = \frac{M_2}{A_2}; \quad E_2 = a_2; \tag{69}$$

$$E_2 = \frac{1.31}{2.3} = 0.57; \quad a_2 = 0.57. \tag{70}$$

The total elasticity of expenses (labour and capital) shows the effect of simultaneous proportional increase in the amount of labor resources and fixed assets:

$$E = E_1 + E_2 = a_1 + a_2 = 0.52 + 0.57 = 1.09. \tag{71}$$

The elasticity of the PF in the point C ($x_0 \cdot y_0$) in modulus is equal to the ratio of distances in the tangent to the point C with coordinates (x_0, $f(x_0)$) to the point of intersection with the axes Y and X (Fig. 5.6).

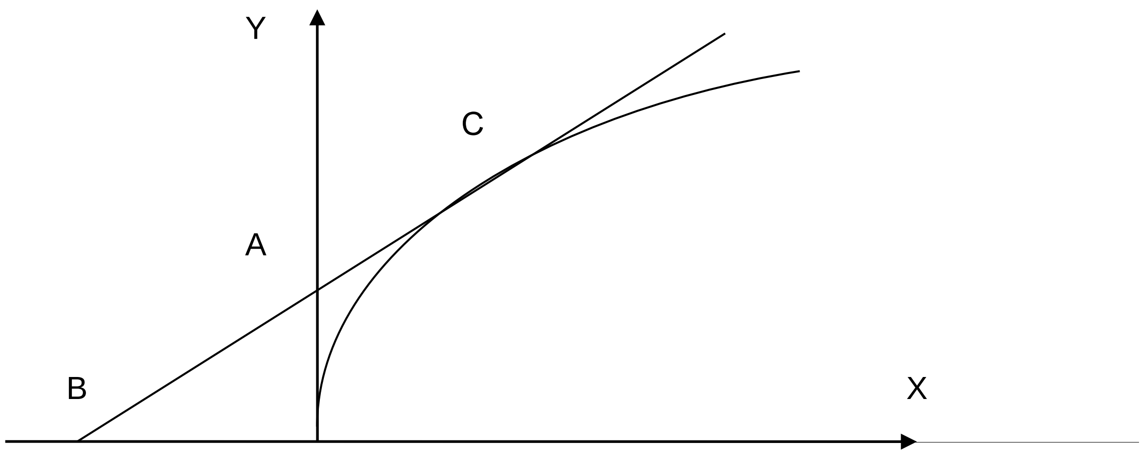


Fig. 5.6. **The elasticity of the production function**

2.5. The capital-labour ratio shows how much of the average capital fits to the unit of spent labor and it is calculated according to the formula:

$$FT = \frac{x_2}{x_1}; \tag{72}$$

$$\frac{x_2}{x_1} = a_0 - \frac{1}{a_2 \cdot y_0} - \frac{1}{a_2 \cdot x_1} - \frac{1 - a_1}{a_2}. \tag{73}$$

$$\frac{100}{100} = 1.52 - \frac{1}{0.57 \cdot 230.06} - \frac{1}{0.57 \cdot 100} - \frac{1 - 0.52}{0.57} = 1. \tag{74}$$

It should be noted that if the sum of indicators in the PF of Cobb – Douglas $Y = a_0 \cdot X_1^{a_1} \cdot X_2^{a_2}$ is equal to one, ($a_1 + a_2 = 1$), then we have:

$$\frac{Y}{x_1} = \frac{a_0 \cdot x_1^{a_1} \cdot x_2^{a_2}}{x_1} = \frac{a_0 \cdot x_2^{a_2}}{x_1^{1-a_1}} = a_0 \cdot \left(\frac{x_2}{x_1}\right)^{a_2}, \tag{75}$$

or if we switch to new designations

$$z = \frac{Y}{x_1}, \tag{76}$$

$$k = \frac{x_2}{x_1}, \tag{77}$$

then we get the following dependence:

$$Z = \mathrm{a}_0 \cdot \mathrm{K}^{\mathrm{a}_2}. \tag{78}$$

Since $0 < a_2 < 1$, the formula implies that the productivity of labor Z is growing more slowly than the capital power.

3. Constructing the isoquants of the production function. The production function allows us to calculate the need for one resource at a given volume of production and the value of another resource.

The need for labor costs with known values of output and capital costs is calculated as:

$$X_1 = \left(\frac{\hat{Y}}{a_0 \cdot x_2^{a_2}}\right)^{\frac{1}{a_1}}. \tag{79}$$

The need for capital costs with known values of output and labor costs is calculated as:

$$X_2 = \left(\frac{\hat{Y}}{a_0 \cdot x_1^{a_1}}\right)^{\frac{1}{a_2}}. \tag{80}$$

The calculation of the need for one of the resources is necessary for the construction of an isoquant of the production function.

The isoquant of the PF is a line of the level $q = f(X_1, X_2)$, $(q > 0)$, that representing a set of points in which the PF takes the value equal to q. The isoquants represent different sets (the ratio) of used resources that provide the same amount of the production output. The graph of the production function isoquant is shown in Fig. 5.7.

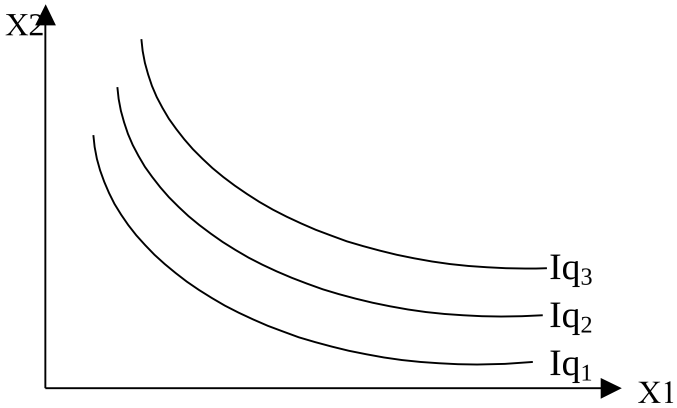


Fig. 5.7. **The isoquants of the production function**

To construct an isoquant of the PF for the production volume Y = 800, we lave to change the value of the amount of the spent capital (Table 5.3) to calculate the need for labor costs and get the following combinations:

$$X_1 = \left(\frac{800}{1.52 \cdot x_2^{0.57}} \right)^{\frac{1}{0.52}}. \quad (81)$$

Table 5.3

The need for labor costs

X_1	100	200	300	400	500	600	700	800	900
X_2	1099	514	330	240	188	154	130	112	99

For a given isoquant, construct an isoclinal of the PF, i.e. a line that connects the origin of the coordinates and the points on the isoquants of the PF for which the marginal rates of substitution of resources will be equal (Fig. 5.8).

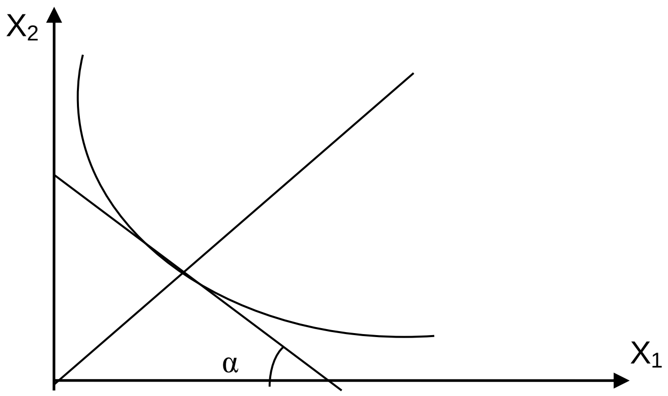


Fig. 5.8. **The isoquant of the PF and the isoclinal**

The limit rate of replacement of the *i*-th resource by the *j*-th resource is calculated as follows:

$$R_{ij} = -\frac{\Delta x_j}{\Delta x_i}, \tag{82}$$

$$R'_{ij} = -\frac{\partial x_j}{\partial x_i}. \tag{83}$$

The marginal rate of replacement of resources R_{ij} shows by how many units the costs of the resource *j* (with a constant fixed output), will increase if the costs of the *i*-th resource decreases by one unit.

For a two-factor production function, the following equality holds:

$$|R_{12}| = \frac{E_1}{E_2} \cdot \frac{x_2}{x_1}. \tag{84}$$

$$|R_{12}| = \frac{0.52}{0.57} \cdot \frac{240}{400} = 0.54. \tag{85}$$

The marginal rate of replacement of the resource coincides with the tangent of the angle of inclination *f* to the axis of the tangent to the isoquant VF in the point (x_0, $f(x_0)$) $R_{12} = \mathrm{tg}(\alpha)$.

The production function can be used to calculate the elasticity of the substitution of factors (resources):

$$\sigma_{ij} = \frac{\partial\left(\frac{x_j}{x_i}\right)}{\frac{x_j}{x_i}} : \frac{\partial R_{ij}}{R_{ij}} \tag{86}$$

The elasticity of substitution of resources has the following economic content: it approximately shows how much the percentage of resources should change (with the fixed output), with the marginal replacement rate R_{ij} change by 1 %.

Practical activity 6. Dynamic econometric models

The goal is to consolidate the theoretical material and to acquire the skills in modeling and analysis of dynamic econometric models for the research in real economic processes.

The following data on the dynamics of sales volume are given in Table 6.1.

Table 6.1

The input data

No.	Date	Sales	No.	Date	Sales
1	November 19, 2014	127.96	15	December 7, 2014	131.7
2	November 20, 2014	127.9	16	December 10, 2014	132.49
3	November 21, 2014	128.37	17	December 11, 2014	132.45
4	November 22, 2014	130.1	18	December 12, 2014	131.72
5	November 23, 2014	129.66	19	December 13, 2014	131.38
6	November 26, 2014	128.79	20	December 14, 2014	132.37
7	November 27, 2014	129.83	21	December 17, 2014	133.51
8	November 28, 2014	130.23	22	December 18, 2014	132.53
9	November 29, 2014	130.22	23	December 19, 2014	133.13
10	November 30, 2014	129.66	24	December 20, 2014	131.92
11	December 3, 2014	129.52	25	December 21, 2014	131.39
12	December 4, 2014	130.43	26	December 24, 2014	131.15
13	December 5, 2014	130.58	27	December 25, 2014	130.95
14	December 6, 2014	131.54	28	December 26, 2014	129.38

According to these data:

1. Build a graph of the dynamics of the indicator and analyse the behaviour of change in its values.

2. Test the availability of a trend in the variance and the mean with Fisher's test, the method of comparison of means and the method of Foster – Stewart.

3. Suggest the hypotheses about the type of trend. Build graphs and estimate the parameters of each trend.

4. Calculate the predicted values of the indicator two steps forward using the model of the time series trend.

5. Provide quality estimates of various models of the trend (the mean error, the mean absolute error, the standard deviation of errors, the mean percentage error, the mean absolute percentage error). Make comparative analysis of the models and determine the most adequate one among them.

6. Provide economic interpretation of the results.

Guidelines

Starting Microsoft Excel and preparing data. Select MS Excel in the program menu. After its launch, enter the input data, as shown in Fig. 6.1.

1. Building the graph. For plotting, select C1:C29 cells with the input data along with the "Sales" headline and select *Plot* (*Plot with Markers*) in the menu item *Insert*. The result is shown in Fig. 6.2.

2. Testing the trend availability. The preliminary stage of selection of the trend of the time series data is testing the hypothesis about the trend availability in the tested process. The most reliable results can be obtained by applying the Fischer's criterion, the Student's criterion and the method of Foster – Stewart.

The Fisher's test is used to determine the trend in the variance.

The input time series $y_1, y_2, \dots, y_n$ is split into two volumes of n_1 and n_2 $\left(n_1 \approx n/2,\ n_2 = n - n_1\right)$:

$$y_1, y_2, \dots, y_k, \tag{87}$$

$$y_{k+1}, y_{k+2}, \dots, y_n. \tag{88}$$

In this case: $n_1 = \dfrac{28}{2} = 14, \quad n_2 = 28 - 14 = 14.$

The result is shown in Fig. 6.3.

	B	C	D
2		Date	Sales
3	1	November 19, 2014	127,96
4	2	November 20, 2014	127,9
5	3	November 21, 2014	128,37
6	4	November 22, 2014	130,1
7	5	November 23, 2014	129,66
8	6	November 26, 2014	128,79
9	7	November 27, 2014	129,83
10	8	November 28, 2014	130,23
11	9	November 29, 2014	130,22
12	10	November 30, 2014	129,66
13	11	December 3, 2014	129,52
14	12	December 4, 2014	130,43
15	13	December 5, 2014	130,58
16	14	December 6, 2014	131,54
17	15	December 7, 2014	131,7
18	16	December 10, 2014	132,49
19	17	December 11, 2014	132,45
20	18	December 12, 2014	131,72
21	19	December 13, 2014	131,38
22	20	December 14, 2014	132,37
23	21	December 17, 2014	133,51
24	22	December 18, 2014	132,53
25	23	December 19, 2014	133,13
26	24	December 20, 2014	131,92
27	25	December 21, 2014	131,39
28	26	December 24, 2014	131,15
29	27	December 25, 2014	130,95
30	28	December 26, 2014	129,38

Fig. 6.1. **The input data**

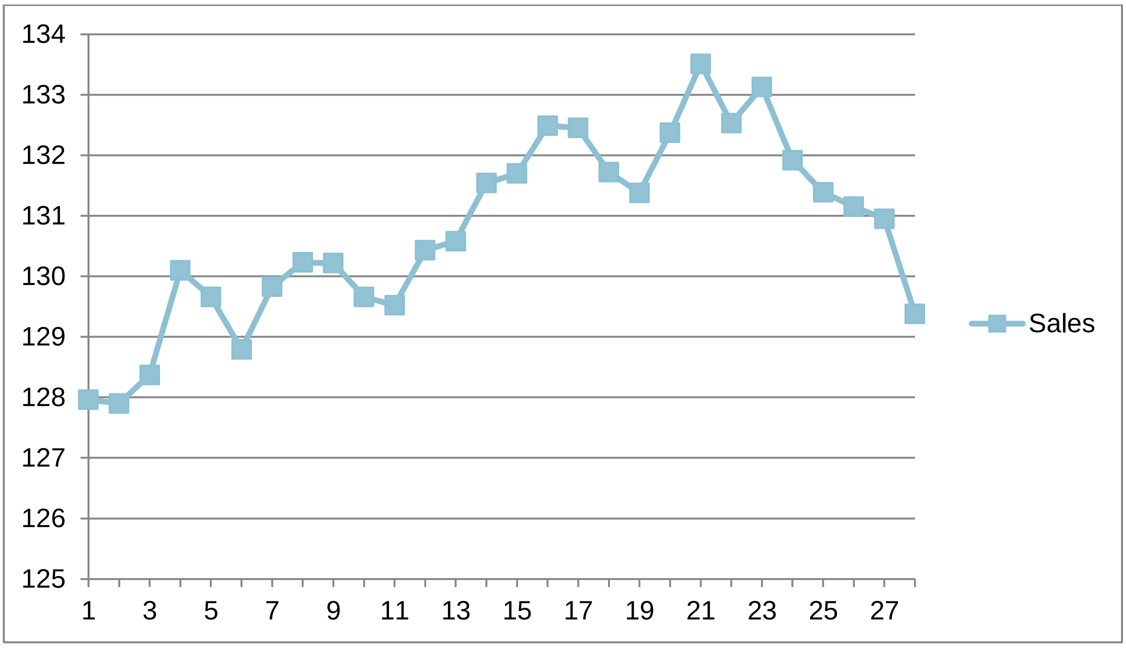


Fig. 6.2. **The input data plot**

D	E	F
	Group 1	Group 2
1	127,96	131,70
2	127,90	132,49
3	128,37	132,45
4	130,10	131,72
5	129,66	131,38
6	128,79	132,37
7	129,83	133,51
8	130,23	132,53
9	130,22	133,13
10	129,66	131,92
11	129,52	131,39
12	130,43	131,15
13	130,58	130,95
14	131,54	129,38

Fig. 6.3. **The result of splitting the input data into two sets**

For each of the sets, the mean and the variance are defined, as shown in Fig. 6.4.

	B	D	E	F
16				
17		Average	=AVERAGE(E2:E16)	=AVERAGE(F2:F16)
18		Variance	=VAR(E2:E16)	=VAR(F2:F16)
19		n_i	=COUNT(E2:E16)	=COUNT(F2:F16)
20		$n_i -1$	=E19-1	=F19-1

Fig. 6.4. **The formulas for calculation**

Fig. 6.5 shows the results of calculating the average values and variances for each set of data.

D	E	F	G
Average	129,6	131,9	
Variance	1,09603	1,05939	
n_i	14	14	
$n_i -1$	13	13	

Fig. 6.5. **The results of the calculation**

The estimated value of the Fisher's test is determined by the formula:

$$F_{theor} = \begin{cases} \dfrac{S_2^2}{S_1^2}, & if \quad S_2^2 > S_1^2 \\ \dfrac{S_1^2}{S_2^2}, & if \quad S_1^2 > S_2^2 \end{cases}. \tag{89}$$

In this case $S_1^2 > S_2^2$, thus $F_{calc} = E18/F18$.

This value is compared with the tabular one for the significance level α = 0.05, and the degrees of freedom $k_1 = n_1 - 1$ and $k_2 = n_2 - 1$ (where k_1 corresponds to the bigger variance). Tabular values can be calculated by the formula: FINV(0,05;E20;F20). The results of the calculation are shown in Fig. 6.6.

	A	B	C	D	E	F	G	H
1		Date	Sales		Group 1	Group 2		
2	1	November 19, 2014	127,96	1	127,96	131,7		
3	2	November 20, 2014	127,9	2	127,9	132,49		
4	3	November 21, 2014	128,37	3	128,37	132,45		
5	4	November 22, 2014	130,1	4	130,1	131,72		
6	5	November 23, 2014	129,66	5	129,66	131,38	F =	1,03459
7	6	November 26, 2014	128,79	6	128,79	132,37	F_{tab} =	2,5769
8	7	November 27, 2014	129,83	7	129,83	133,51		
9	8	November 28, 2014	130,23	8	130,23	132,53		
10	9	November 29, 2014	130,22	9	130,22	133,13		
11	10	November 30, 2014	129,66	10	129,66	131,92		
12	11	December 3, 2014	129,52	11	129,52	131,39		
13	12	December 4, 2014	130,43	12	130,43	131,15		
14	13	December 5, 2014	130,58	13	130,58	130,95		
15	14	December 6, 2014	131,54	14	131,54	129,38		
16	15	December 7, 2014	131,7					
17	16	December 10, 2014	132,49	Average	129,6	131,9		
18	17	December 11, 2014	132,45	Variance	1,09603	1,05939		
19	18	December 12, 2014	131,72	n_i	14	14		
20	19	December 13, 2014	131,38	$n_i - 1$	13	13		

Fig. 6.6. **The results of the calculation**

If $F_{calc} \geq F_{tabl}$ (α, k_1, k_2), then the hypothesis about the trend in the variance is confirmed. In this case, $F_{calc} < F_{tabl}$ (α, k_1, k_2), thus the hypothesis about the trend in the variance is not confirmed, and it can be assumed that this trend in the time series is missing.

After analyzing the availability of a trend in the variance, we proceed to the analysis of the availability of a trend in the mean value using the method of mean comparison.

For this purpose, we calculate the value of t_{calc} according to the formula:

$$t_{calc} = \frac{\left|\overline{y}_1 - \overline{y}_2\right|}{\sqrt{(n_1 - 1) \cdot S_1^2 + (n_2 - 1) \cdot S_2^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}. \tag{90}$$

The formulas for calculation are shown in Fig. 6.7.

	F	G	H
1	Group 2		
2	131,7		
3	132,49		
4	132,45		
5	131,72		
6	131,38	F =	=E18/F18
7	132,37	F_{tab} =	=FINV(0,05;E20;F20)
8	133,51	t =	=ABS(E17-F17)/SQRT(E20*E18+F20*F18)*SQRT((E19*F19*(E19+F19-2))/(E19+F19))
9	132,53	t_{tab} =	=TINV(0,05;E19+F19-2)
10	133,13		
11	131,92		
12	131,39		
13	131,15		
14	130,95		
15	129,38		
16			
17	=AVERAGE(F2:F16)		
18	=VAR(F2:F16)		
19	=COUNT(F2:F16)		
20	=F19-1		

Fig. 6.7. **The formulas for calculation**

The estimated value of the Student's t-test is compared with the tabular one (t_{tabl}) with a confidence level of α and the number of degrees of freedom $k = n - 2$. In this case, $t_{calc} > t_{tabl}$, then the hypothesis about the trend in the mean value is confirmed. And, as $\overline{y_2} = \overline{y_1}$, the trend is ascending.

The results of the calculation are shown in Fig. 6.8.

Let's consider the use of the method of Foster – Stewart for the same task.

The value of the level y_1 is called the record one if it is more than any of the previous ones or less than all the previous values.

	D	E	F	G	H
1		Group 1	Group 2		
2	1	127,96	131,7		
3	2	127,9	132,49		
4	3	128,37	132,45		
5	4	130,1	131,72		
6	5	129,66	131,38	F =	1.03459
7	6	128,79	132,37	F_{tab} =	2.57691
8	7	129,83	133,51	t =	5.69425
9	8	130,23	132,53	t_{tab} =	2.05553
10	9	130,22	133,13		
11	10	129,66	131,92		
12	11	129,52	131,39		
13	12	130,43	131,15		
14	13	130,58	130,95		
15	14	131,54	129,38		
16					
17	Average	129,6	131,9		
18	Variance	1,09603	1,05939		
19	n_i	14	14		
20	$n_i - 1$	13	13		

Fig. 6.8. **The results of the calculation**

The values u_i and v_i are calculated by the formulas:

$$u_i = \begin{cases} 1 & \text{if } y_i \text{ is less than the previous ones,} \\ 0 & \text{otherwise.} \end{cases} \quad (91)$$

$$v_i = \begin{cases} 1 & \text{if } y_i \text{ is less than the previous ones,} \\ 0 & \text{otherwise.} \end{cases} \quad (92)$$

To do this, we create a table header in cells I1, J1, K1, L1. In line with the number 2 dashes should be put.

In the third line, the formula is entered into cell I3: =IF(C3=MAX($C$2:C3);1;0). In this formula, only the beginning of the range is "fixed". Then the formula is extended down to the 28th element.

In the next column J in cell J3, a similar formula (=IF(C3=MIN($C$2:C3);1;0) is entered and extended down.

The formulas for calculating are shown in Fig. 6.9.

	I	J	K	L
1	u_i	v_i	s_i	d_i
2				
3	=IF(C3=MAX($C$2:C3);1;0)	=IF(C3=MIN($C$2:C3);1;0)	=I3+J3	=I3-J3
4	=IF(C4=MAX($C$2:C4);1;0)	=IF(C4=MIN($C$2:C4);1;0)	=I4+J4	=I4-J4
5	=IF(C5=MAX($C$2:C5);1;0)	=IF(C5=MIN($C$2:C5);1;0)	=I5+J5	=I5-J5
6	=IF(C6=MAX($C$2:C6);1;0)	=IF(C6=MIN($C$2:C6);1;0)	=I6+J6	=I6-J6
7	=IF(C7=MAX($C$2:C7);1;0)	=IF(C7=MIN($C$2:C7);1;0)	=I7+J7	=I7-J7
8	=IF(C8=MAX($C$2:C8);1;0)	=IF(C8=MIN($C$2:C8);1;0)	=I8+J8	=I8-J8
9	=IF(C9=MAX($C$2:C9);1;0)	=IF(C9=MIN($C$2:C9);1;0)	=I9+J9	=I9-J9
10	=IF(C10=MAX($C$2:C10);1;0)	=IF(C10=MIN($C$2:C10);1;0)	=I10+J10	=I10-J10
11	=IF(C11=MAX($C$2:C11);1;0)	=IF(C11=MIN($C$2:C11);1;0)	=I11+J11	=I11-J11
12	=IF(C12=MAX($C$2:C12);1;0)	=IF(C12=MIN($C$2:C12);1;0)	=I12+J12	=I12-J12
13	=IF(C13=MAX($C$2:C13);1;0)	=IF(C13=MIN($C$2:C13);1;0)	=I13+J13	=I13-J13
14	=IF(C14=MAX($C$2:C14);1;0)	=IF(C14=MIN($C$2:C14);1;0)	=I14+J14	=I14-J14
15	=IF(C15=MAX($C$2:C15);1;0)	=IF(C15=MIN($C$2:C15);1;0)	=I15+J15	=I15-J15
16	=IF(C16=MAX($C$2:C16);1;0)	=IF(C16=MIN($C$2:C16);1;0)	=I16+J16	=I16-J16
17	=IF(C17=MAX($C$2:C17);1;0)	=IF(C17=MIN($C$2:C17);1;0)	=I17+J17	=I17-J17
18	=IF(C18=MAX($C$2:C18);1;0)	=IF(C18=MIN($C$2:C18);1;0)	=I18+J18	=I18-J18
19	=IF(C19=MAX($C$2:C19);1;0)	=IF(C19=MIN($C$2:C19);1;0)	=I19+J19	=I19-J19
20	=IF(C20=MAX($C$2:C20);1;0)	=IF(C20=MIN($C$2:C20);1;0)	=I20+J20	=I20-J20
21	=IF(C21=MAX($C$2:C21);1;0)	=IF(C21=MIN($C$2:C21);1;0)	=I21+J21	=I21-J21
22	=IF(C22=MAX($C$2:C22);1;0)	=IF(C22=MIN($C$2:C22);1;0)	=I22+J22	=I22-J22
23	=IF(C23=MAX($C$2:C23);1;0)	=IF(C23=MIN($C$2:C23);1;0)	=I23+J23	=I23-J23
24	=IF(C24=MAX($C$2:C24);1;0)	=IF(C24=MIN($C$2:C24);1;0)	=I24+J24	=I24-J24
25	=IF(C25=MAX($C$2:C25);1;0)	=IF(C25=MIN($C$2:C25);1;0)	=I25+J25	=I25-J25
26	=IF(C26=MAX($C$2:C26);1;0)	=IF(C26=MIN($C$2:C26);1;0)	=I26+J26	=I26-J26
27	=IF(C27=MAX($C$2:C27);1;0)	=IF(C27=MIN($C$2:C27);1;0)	=I27+J27	=I27-J27
28	=IF(C28=MAX($C$2:C28);1;0)	=IF(C28=MIN($C$2:C28);1;0)	=I28+J28	=I28-J28
29	=IF(C29=MAX($C$2:C29);1;0)	=IF(C29=MIN($C$2:C29);1;0)	=I29+J29	=I29-J29
30			=SUM(K3:K29)	=SUM(L3:L29)

Fig. 6.9. **The formulas for calculation**

Next, the values $s_i = u_i + v_i$ and $d_i = u_i - v_i$, are calculated as well as:

$$S = \sum_{i=2}^{n} s_i, \quad D = \sum_{i=2}^{n} d_i. \tag{93}$$

For each of these indicators, the value of the Student's t-test is calculated according to the formulas:

$$t_D = \frac{|D|}{\sigma_D(n)}, \quad t_S = \frac{S - \mu(n)}{\sigma_S(n)}, \tag{94}$$

where $\mu(n)$, $\sigma_S(n)$, $\sigma_D(n)$ are table values.

The calculated values of the Student's t-test are compared with the tabular ones with the confidence level *α* and the number of degrees of freedom $k = n - 2$.

While using the method of Foster – Stewart the following situations may occur:

a) $t_D > t_p$, $t_S > t_p$ – the hypothesis about the trend availability in the mean value is accepted, while if $D > 0$, the trend is ascending, and if $D < 0$, the trend is descending (in this case, the series monotonically ascends or descends);

b) $t_D < t_p$, $t_S > t_p$ – the hypothesis about the trend availability in the variance is accepted (in this case, the fluctuations of the indicator occur);

c) $t_D > t_p$, $t_S < t_p$ – the hypothesis about the trend availability in the variance or in the mean value cannot be accepted or rejected;

d) $t_D < t_p$, $t_S < t_p$ – the hypothesis about the trend absence both in the variance and the mean value can be accepted.

3. Selection of the type of trend.

The data line on the plot should be clicked on with the right mouse button and *Add trend line…* selected as shown in Fig. 6.10.

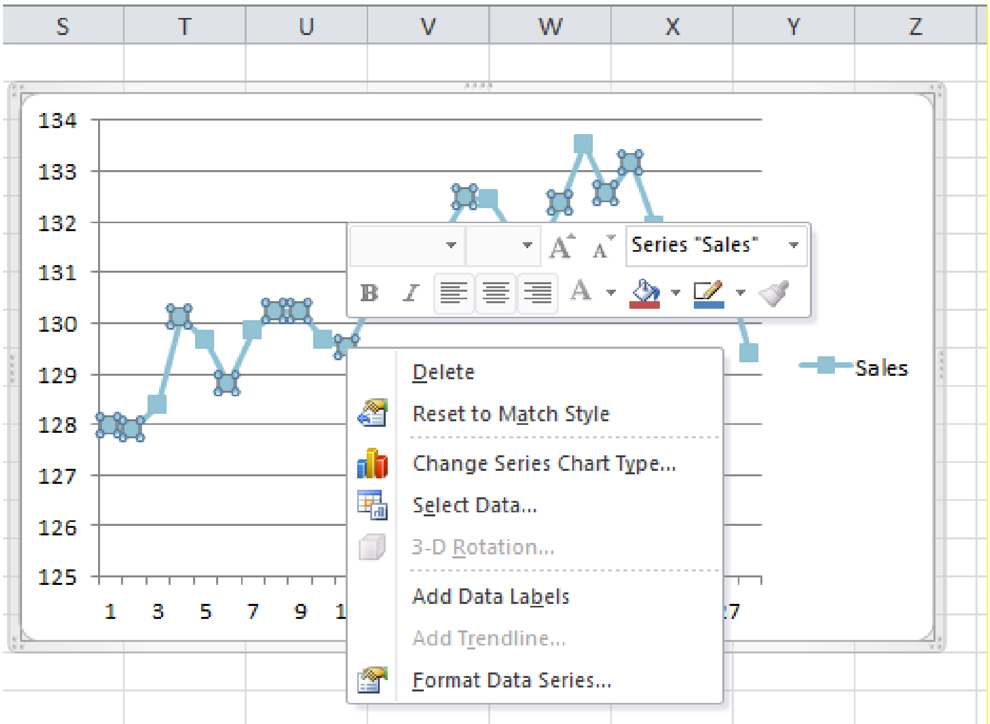


Fig. 6.10. **Adding the trend line**

It is necessary to choose the type of the trend line (e.g., *Linear*), set the forecast horizon (e.g., *2 periods forward*), tick *show the equation on the plot*, and *put the value of approximation reliability on the plot* as shown in Fig. 6.11.

Then consequently try alternative types of trend.

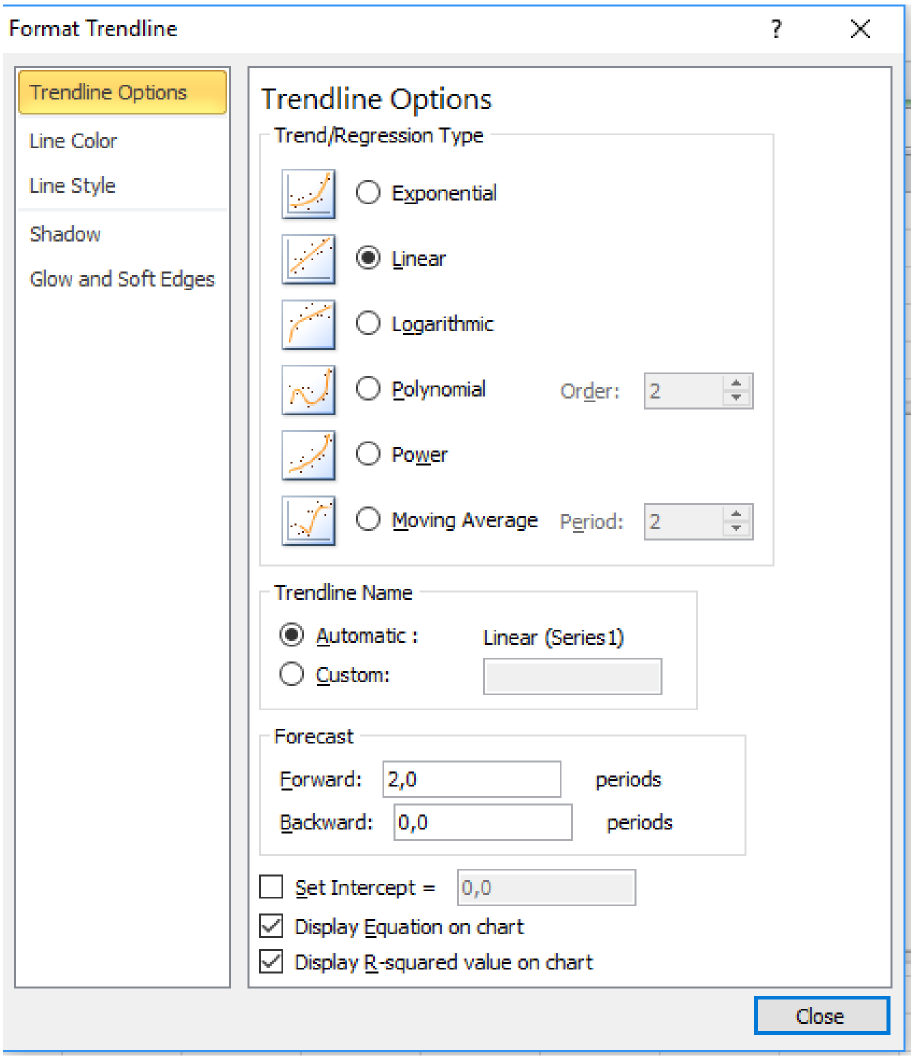


Fig. 6.11. **Selection of parameters of the trend line**

The results of building various trends are shown in Fig. 6.12 – 6.17.

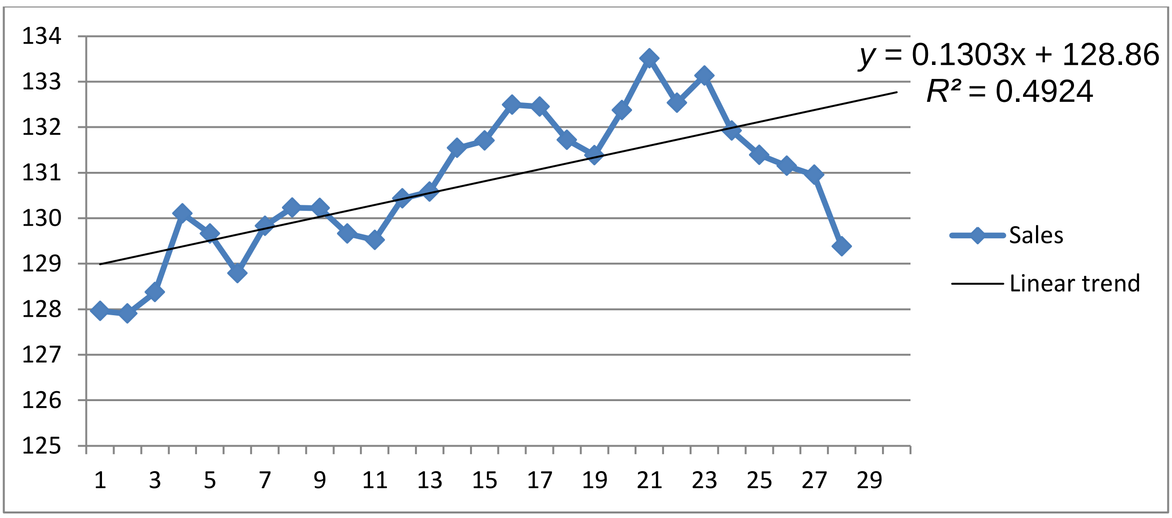


Fig. 6.12. **The results of building the linear trend**

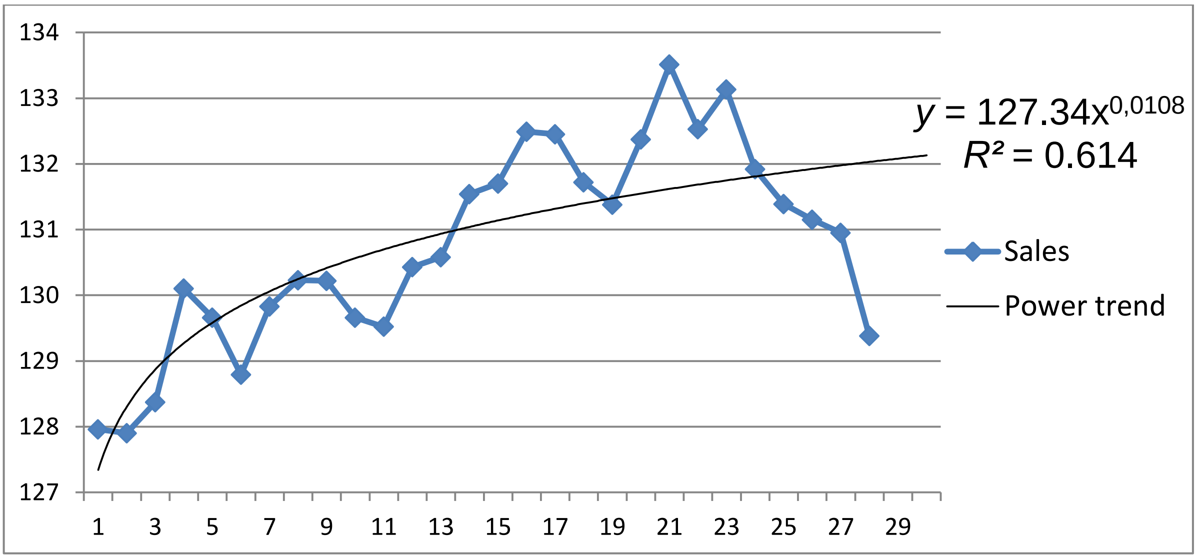


Fig. 6.13. **The results of building the exponential trend**

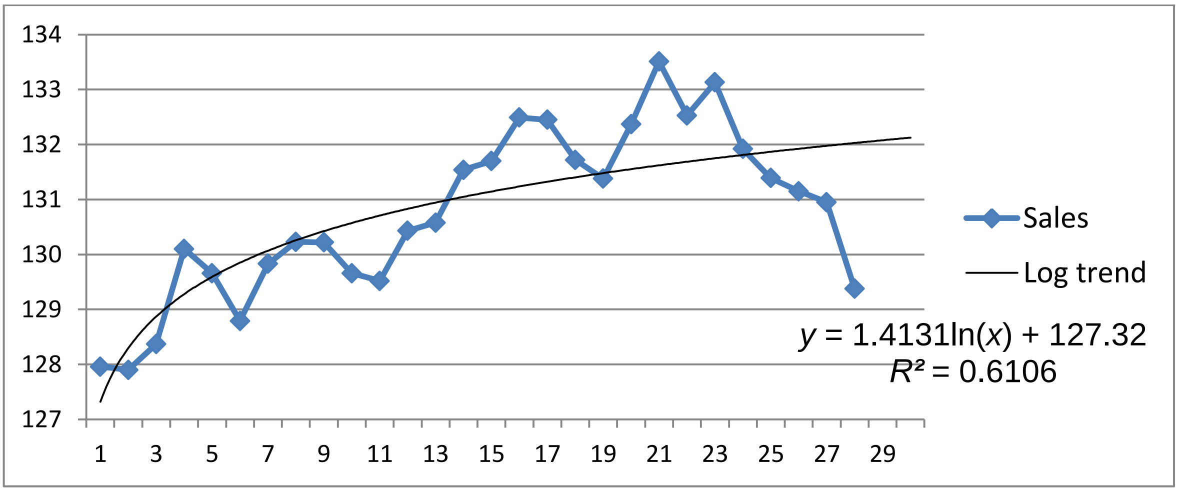


Fig. 6.14. **The results of building the logarithmic trend**

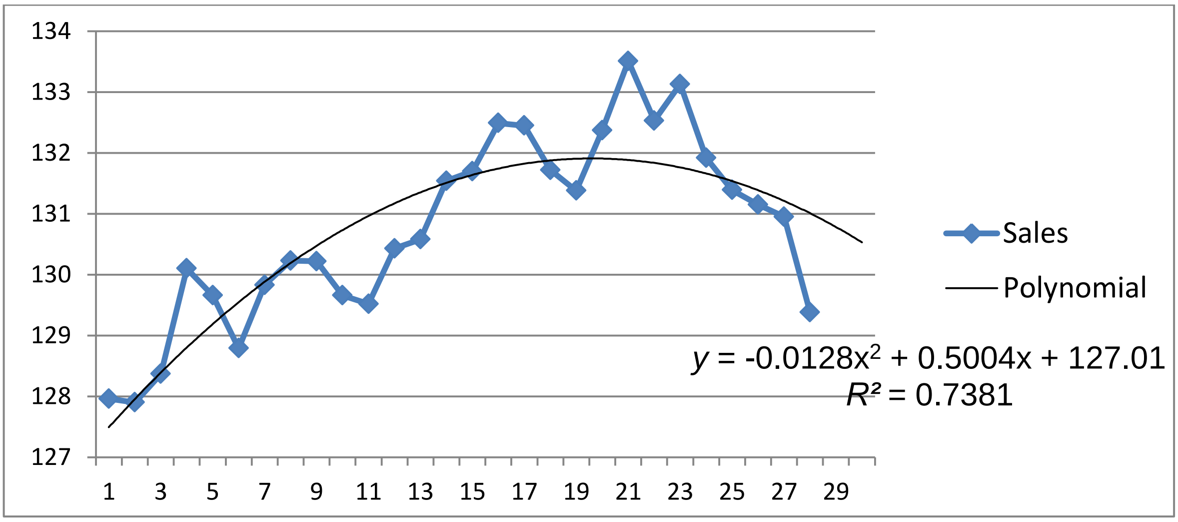


Fig. 6.15. **The results of building the polynomial trend of the 2nd degree**

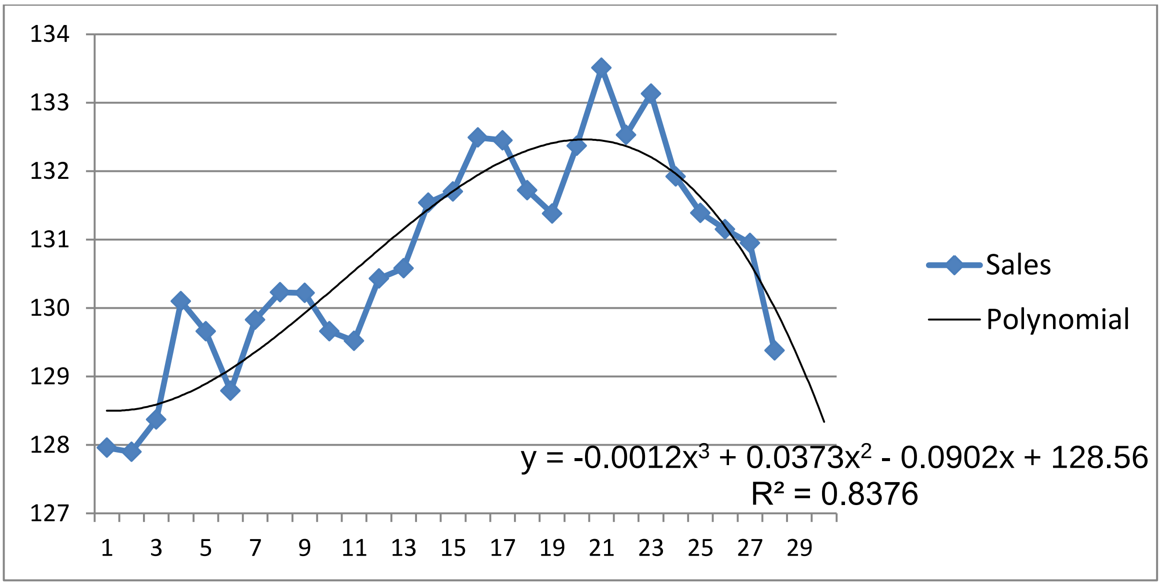


Fig. 6.16. **The results of building the polynomial trend of the 3rd degree**

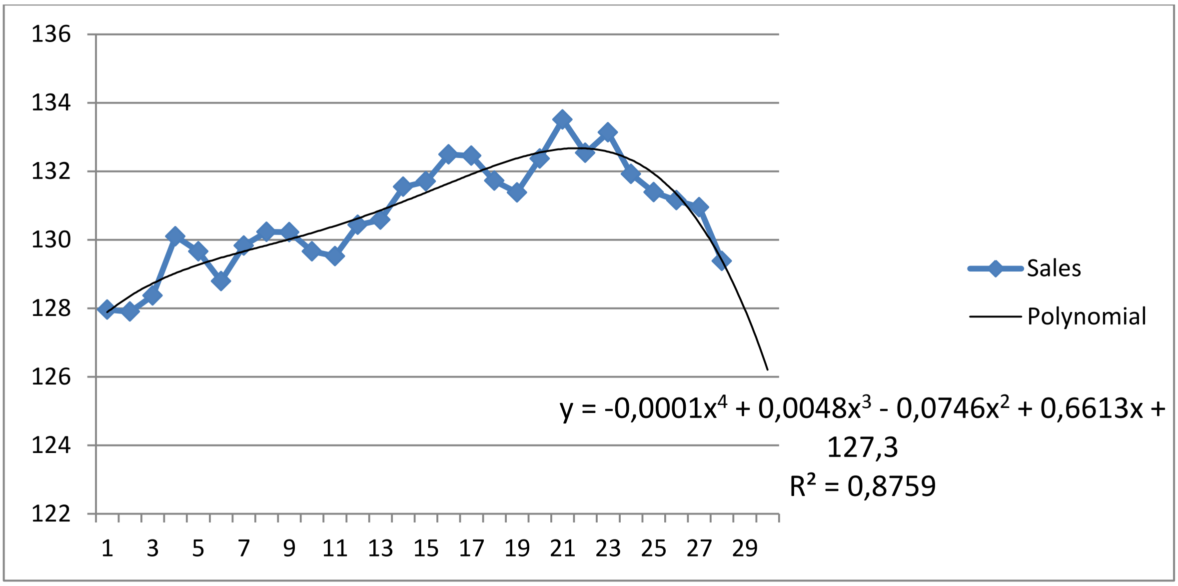


Fig. 6.17. **The results of building the polynomial trend of the 4th degree**

4. Forecasting.

According to the given trend equations, the theoretic values of the sales volume should be calculated (*Predict* ($\hat{y}$)).

5. Quality estimates of various models of the trend.

To explore such quality estimates of the time series model as the mean error, the mean absolute error, the standard deviation of errors, the mean percentage error, the mean absolute percentage error, a new *Residuals* (*Model errors*) variable should be entered and calculated by setting the formula *Residuals = Sales Volume – Predict*.

The formulas for calculating the estimates of the model accuracy are shown in Table 6.2.

Table 6.2

The characteristics of the estimation of the model accuracy

Name	Formula
Mean error	$m.e. = \dfrac{\sum_{t=1}^{n} e_t}{n}$
Mean absolute error	$m.a.e. = \dfrac{\sum_{t=1}^{n} \lvert e_t \rvert}{n}$
Sum of squares of errors	$s.s.e. = \sum_{t=1}^{n} e_t^2$
Mean squared error	$m.s.e. = \sqrt{\dfrac{\sum_{t=1}^{n} e_t^2}{n}}$
Mean percentage error	$m.p.e. = \dfrac{1}{n} \sum_{t=1}^{n} \dfrac{e_t}{y_t} \cdot 100\ \%$
Mean absolute percentage error	$m.a.p.e. = \dfrac{1}{n} \sum_{t=1}^{n} \dfrac{\lvert e_t \rvert}{y_t} \cdot 100\ \%$

Draw a comparison of the models for the value of the average absolute percentage error. If the value of the average absolute percentage error is in the range:

0 < m.a.p.e < 10 %, the model provides a high prediction accuracy;

10 % < m.a.p.e < 20 %, the model provides a satisfactory accuracy of the forecast;

m.a.p.e > 20 %, the model is not adequate.

Recommended literature

11.1. Basic

1. Боровиков В. П. Популярное введение в программу STATISTICA / В. П. Боровиков. – Москва : Компьютер Пресс,1998. – 194 с.

2. Доугерти К. Введение в эконометрику / К. Доугерти ; пер. с англ. – Москва : ИНФРА-М, 1997. – 402 с.

3. Економетрика : навчальний посібник / Л. С. Гур'янова, Т. С. Клебанова, О. А. Сергієнко та ін. – Харків : Вид. ХНЕУ ім. С. Кузнеця, 2015. – 389 с.

4. Клебанова Т. С. Эконометрия / Т. С. Клебанова, Н. А. Дубровина, Е. В. Раевнева. – Харьков : ИД "ИНЖЭК", 2003. – 128 с.

5. Наконечний С. І. Економетрія / С. І. Наконечний, Т. О. Терещенко, Т. П. Романюк. – Київ : КНЕУ, 1997. – 352 с.

6. Прогнозування соціально-економічних процесів : навчальний посібник / Т. С. Клебанова, В. А. Курзенев, В. М. Наумов та ін. – Харків : Вид. ХНЕУ ім. С. Кузнеця, 2015. – 656 с.

7. Эконометрия на персональном компьютере / Т. С. Клебанова, Н. А. Дубровина, А. В. Милов и др. – Харьков : Изд. ХГЭУ, 2002. – 208 с.

11.2. Additional

8. Боровиков В. П. STATISTICA: искусство анализа данных на компьютере. Для профессионалов / В. П. Боровиков. – Санкт-Петербург : Питер, 2001. – 656 с.

9. Лук'яненко І. Економетрика / І. Лук'яненко, Л. Краснікова. – Київ : Товариство "Знання", КОО, 1998. – 494 с.

10. Магнус Я. Р. Эконометрика. Начальный курс / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – Москва : Дело, 1997. – 248 с.

11. Орлов. А. Н. Эконометрика / А. Н. Орлов. – Москва : Изд. "Экзамен", 2002. – 576 с.

12. Черняк О. І. Динамічна економетрика / О. І. Черняк, А. В. Ставицький. – Київ : КВІЦ, 2000. – 120 с.

13. Hansen Bruce E. Econometrics / Bruce E. Hansen. – Wisconsin : University of Wisconsin, 2016. – 381 p.

14. Hubler O. Modern Econometric Analysis: Surveys on Recent Developments / O. Hubler, J. Frohn. – s. l. : Springer, 2006. – 234 p.

15. Verbeek M. A Guide to Modern Econometrics / M. Verbeek. – 2nd ed. – S. l. : Wiley, 2004. – 446 p.

16. Wooldridge J. M. Introductory Econometrics : A Modern Approach / Jeffry M. Wooldridge. – 4th edition. – Thompson : South-Western College Publishing, 2012. – 912 p.

11.3. Information resources

17. Економетрика – бібліотека ресурсів [Електронний ресурс]. – Режим доступу : http://efaculty.kiev.ua/ekon.htm.

18. Статистика України : науковий журнал [Електронний ресурс]. – Режим доступу : www.ukrstat.gov.ua.

19. Сайт Агентства по развитию инфраструктуры фондового рынка Украины [Электронный ресурс]. – Режим доступа : http://www.smida.gov.ua/db.

20. Сайт Національного банку України – Режим доступу : www.bank.gov.ua.

21. Сайт ПФТС – Режим доступу : http://pfts.com.

11.4. Methodical support

22. Guryanova L. Guidelines to laboratory sessions of the academic discipline "Mathematical Modeling in Economics and Management: Econometrics" [Electronic resource] / L. Guryanova, O. Sergienko, S. Prokopovych. – Access mode : http://elearn2.ekhneu.org.ua/main/document/document.php?cidReq=MMMECONOMETR&id_session=0&gidReq=0&origin=.

23. Sergienko O. Tests on the academic discipline "Mathematical Modeling in Economics and Management: Econometrics" [Electronic resource] / O. Sergienko, S. Milevskyi, S. Prokopovych. – Access mode : http://elearn2.ekhneu.org.ua/main/exercice.

24. Prokopovych S. Mathematical modeling in economics and management: Econometrics : reference compendium [Electronic resource] / S. Prokopovych, O. Sergienko, S. Milevskyi. – Access mode : http://elearn2.ekhneu.org.ua/main/document/document.php?cidReq=MMMECONOMETR&id_session=0&gidReq=0&origin=&id=53.

Content

НАВЧАЛЬНЕ ВИДАННЯ

ЕКОНОМЕТРИКА

Практикум
для студентів усіх спеціальностей
першого (бакалаврського) рівня

(англ. мовою)

Самостійне електронне текстове мережеве видання

Укладачі: **Гур'янова** Лідія Семенівна
Прокопович Світлана Валеріївна
Мілевський Станіслав Валерійович

Відповідальний за видання *Т. С. Клебанова*

Редактор *З. В. Зобова*

Коректор *З. В. Зобова*

Розглянуто основні питання аналізу та прогнозування соціально-економічних і фінансових процесів і систем на основі застосування економетричних методів і моделей. Наведено практикум з навчальної дисципліни за допомогою програми Microsoft Excel.

Рекомендовано для студентів усіх спеціальностей.

План 2018 р. Поз. № 297 ЕВ. Обсяг 82 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.