УДК 519.862.6
ББК 65в63I
Э 40

Рекомендовано к изданию научно-методическим советом учреждения образования “Белорусский торгово-экономический университет потребительской кооперации”. Протокол № 2 от 13 декабря 2016 г.

Авторы-составители: Л. П. Авдашкова, канд. физ.-мат. наук, доцент;
М. А. Грибовская, канд. физ.-мат. наук, доцент;
С. В. Кравченко, канд. физ.-мат. наук, доцент

Written and compiled by: L. Avdashkova, Candidate of Physico-mathematical sciences, Associate Professor;
M. Gribovskaya, Candidate of Physico-mathematical sciences, Associate Professor;
S. Kravtchenko, Candidate of Physico-mathematical sciences, Associate Professor

Рецензенты: Е. В. Коробейникова, канд. физ.-мат. наук, заместитель начальника отдела внедрения и поддержки ERP-решений производственного управления “Связьинформсервис” РУП ПО “Белоруснефть”;
О. И. Еськова, канд. техн. наук, доцент Белорусского торгово-экономического университета потребительской кооперации

Reviewers: E. Korobeinikova, Candidate of Physico-mathematical sciences, head of Sector of maintenance and support of the production management of ERP solutions of the Department of Communication and Information Technologies “Sviazinformservis” of corporation “Belorusneft”;
A. Yaskova, Candidate of Engineering sciences, Associate Professor Belarusian Trade and Economics University of Customer Cooperatives, Associate Professor

ISBN 978-985-540-451-5

EXPLANATORY NOTE

The discipline “Econometrics (advanced)” examines the technique of construction of generalized linear multiple regression models, econometric models based on stationary and non-stationary time series of econometric analysis, which consists in the diagnosis of models, as well as the methodology for developing forecasts based on econometric models.

Learning discipline “Econometrics (advanced)” requires a knowledge of the theory of probability and mathematical statistics, a basic level of econometrics, economic theory and macroeconomic analysis.

Control of knowledge by means of con-controlling works and tests using personal computers-ditch and related software.

Increasingly complex economic processes require raising the level of education of modern experts in economics and management. Learning discipline “Econometrics (advanced level)” allows the use of modeling and quantitative analysis in economic research. The specialist economic profile for the task must master the scientific principles of research sotsi-cially-economic systems, the analysis of the original data, formalization, forecasting and making optimal management decisions, using for this purpose the modern technical means.

Manual “Econometrics (advanced level)” is designed for graduate students to perform classroom and independent work.

The manual covers the basic econometric issues related to the following:

- construction of the classical linear regression vogue-ley;
- time series analysis of the structure and the study of the relationships between them;
- consideration of the main errors that occur when violations SRI-classical model assumptions, the method of diagnosis and elimination;
- the study of econometric models, denominated system of simultaneous equations.

Econometric modeling is to receive a model and its analysis of the quality of certain statistical parameters that must be found by using mathematical statistics knowledge. Initially, therefore, invited to perform a calculation of parameters using MS Excel application, described in the section on computing technology in MS Excel, and then further analysis and evaluation of building models using the obtained values of the parameters. Such an approach in the construction of classes allows, on the one hand, to recall the details of the mathematical statistics and apply them to solve specific problems of econometric modeling at the stage of calculation, on the other hand, at the stage of analysis to focus on the need to perform spe-

cific statistical conditions, without being distracted by performing calculations.

For each topic in the manual is invited to computing technology, analysis of the received parameters corresponding model questions for self-control, individual tasks. Since the steps of econometric modeling are valid for any model in the topics 1–4 on computing technology topics and analysis of the use-pattern forms a same numbering of these stages.

In the section on the implementation of the construction-econometric analysis of the model for each stage of econometric modeling is appropriate theoretical material, which allows you to organize the discussion of problematic situations, to improve the organization of independent work, answer questions for self-control.

Individual jobs allow you to work independently of undergraduates.

INTRODUCTION IN ECONOMETRICS

The term "econometrics" is believed to have been crafted by Ragnar Frisch of Norway, one of the three principal founders of the Econometric Society, first editor of the journal Econometrica, and co-winner of the first Nobel Memorial Prize in Economic Sciences in 1969. It is therefore fitting that we turn to Frisch's own words in the introduction to the first issue of Econometrica to describe the discipline. A word of explanation regarding the term econometrics may be in order. Its definition is implied in the statement of the scope of the Econometric Society, in Section I of the Constitution, which reads: "The Econometric Society is an international society for the advancement of economic theory in its relation to statistics and mathematics.... Its main object shall be to promote studies that aim at a unification of the theoretical quantitative and the empirical-quantitative approach to economic problems...". But there are several aspects of the quantitative approach to economics, and no single one of these aspects, taken by itself, should be confounded with econometrics. Thus, econometrics is by no means the same as economic statistics. Nor is it identical with what we call general economic theory, although a considerable portion of this theory has a defininitely quantitative character. Nor should econometrics be taken as synonomous with the application of mathematics to economics. Experience has shown that each of these three view-points, that of statistics, economic theory and mathematics, is a necessary, but not by itself a sufficient, condition for a real understanding of the quantitative relations in modern economic life. It is the unification of all three that is powerful. And it is this unification that constitutes econometrics. This definition re-

mains valid today, although some terms have evolved somewhat in their usage. Today, we would say that econometrics is the unified study of economic models, mathematical statistics and economic data. Within the field of econometrics there are sub-divisions and specializations. Econometric theory concerns the development of tools and methods and the study of the properties of econometric methods. Applied econometrics is a term describing the development of quantitative economic models and the application of econometric methods to these models using economic data.

Economic studies require the skills of economists at-changing economic and mathematical methods to create econometric-parameter model based on the knowledge of economic theory, economic statistics, mathematical modeling, theory of probability, mathematical statistics, for understanding the quantitative governmental linkages between economic factors in order to analyze and predict real economic processes.

Economic theory using qualitative analysis establishes a set of factors and indicators influencing the economic phenomenon under study, their role and the theoretical relationship. Economic statistics provide the information base of economic research, carrying out the primary processing of empirical values of selected economic indicators. Economic statistics are generally limited to simple qualitative conclusions. Mathematical Statistics provides a tool for working with random variables. Mathematical modeling formalizes examines the economic problem in the language of mathematics. Econometrics estimates the quantitative relationship of the studied factors and the use of these est one of the main tasks is to build econometric analysis and econometric models. Thus, under the econometric model means the preferred form of presentation of the study of economic problems with the help of mathematical terms and relations on the basis of statistical data, which is useful for quantitative analysis.

There are different classifications of econometric models. For example, one of them is a time factor types of econometric models (static and dynamic, i. e. time-series model). The first of these examined the state of the system at a given time, i. e. based on a one-time cross-section of information on the object under study. The latter are based on the data describing the objects under study for a number of consecutive periods, i. e. they use not only the current values of the indicators, but also some previous time values, as well as the very time *t*.

Depending on the form of the mathematical representation, econometric models are divided into models with one equation and a model of a system of simultaneous equations.

In the first case, the explanatory factors *y* expressed by the explanatory variables with a single equation in which each particular set of explanatory

factors corresponds to a probability value of the dependent variable *y*, expressed expectation *M* (*y*). Depending on the type of function econometric models divided into linear and nonlinear.

The system of equations econometric models used in the study of economic phenomena rather complex relationship between the studied parameters, which are not described by one but by several equations (for example, demand equilibrium model and demand in a market economy).

There are the following stages of solving econometric problems:

- Step statement of the problem, involving the definition of the goals and objectives of the study; allocation of factors and parameters that define the studied economic processes; the establishment on the basis of economic theory, the role of selected indicators.
- Step specification, during which the selected connection between the formula variables that indicate the selected factors. This formula has the general form and contains the parameters (coefficients of the variables) requiring statistical evaluation.
- Step parameterization, the decisive problem of estimating the values of the parameters of the selected communication function.
- Step verification, involving checking the adequacy of the model, i. e. check the model fit the actual economic phenomena or processes. In addition, at this stage it becomes clear how successfully solved the problem of specification and parameterization, improved form of the model clarifies the composition of the explanatory variables, set the accuracy of the calculations for this model are determined by the overall quality of the equation, the statistical significance of the parameters found, as well as many other issues are resolved, determining the reliability of the conclusions of the model.

The overall objective of econometric modeling is as follows: according to available data *n* observations for signs of change *y*, depending on the set of factors values select an econometric model of $y = f(x) + e$, to evaluate its options and statistically prove that the factors are important, and constructed a function $f(x)$ is such that most closely matches the observation data.

THEME 1. MULTIPLE REGRESSION

Formulation of the problem

Explore dependent factor y from x_1 and x_2 factors, using observational data presented in the Table 1. Construct a regression model $y = f(x_1, x_2) + \varepsilon$. Calculate the value of the index y for $x_1 = 35$ and $x_2 = 10$.

Table 1 – **Observations**

Factors		
y	x_1	x_2
684,48	28	10
674,45	26	8
729,62	30	14
748,86	35	15
761,44	41	16
773,42	45	17
628,07	27	3
731,84	35	13
698,81	30	10
645,92	23	5
664,64	29	7
711,18	33	11
798,07	40	20
833,82	41	24
667,97	41	6
607,61	23	2
711,76	32	12
728,62	37	13
666,15	31	7
683,70	30	9

Computing technology in MS Excel to building linear multiple regression model

1. Staging step: defining the goals and objectives of the study; allocation of factors and parameters that define the studied economic processes; the establishment of the role of selected indicators (given in the problem, or determined on the basis of economic theory); preparation of data for calculations

Prepare data for calculations (enter source data presented in Table 1). In cell A1, enter the title of the first column – "y" in cell B1 – the name of the

second column – “x1” in cell C1 – the name of the third column – “x2”. The cells A2, A3, ..., A21, enter the data of the first column of Table 1 in cell B2, B3, ..., B21 – data of the second column, the cells C2, C3, ..., C21 – the third column of the table data.

Enter a new sheet name “Regression” and save the workbook (*File* → *Save as* → ...).

Note:

1. Annex A is an example of design calculations in MS Excel sheets.
2. In MS Excel sheets variables x_1, x_2 will be denoted x_1, x_2.

2. Specification: choice in general formula relationships between variables that indicate the selected factors

The type and strength of the functional dependence (linear or nonlinear) are determined by the correlation coefficient. On the menu Data click Data Analysis.

Note – If the menu is not this command, you should select *File* → *Options* → *Add-ins* → *Analisis ToolPak* → *Go* → *Analisis ToolPak* → *OK*.

On a sheet of “Data”, select *Data* → *Data Analysis* → *Correlation* → *OK*. The values of the window, set the parameters as follows:

- *Input Range* – enter a reference to cell A1:C21;
- *Labels in First Row* – check the box;
- *Output Options* – set the switch to the Output Range and in the field, set a reference to the cell E2. Click OK. Copy the cell in the G5 H4. In cell E1, type the name of the “Correlation matrix”.

To test the hypothesis on the significance of the correlation coefficient compared the observed $t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ and the critical $t_{a,v} = t_{cr}$ value of the Student’s statistics, to find that, follow these steps:

- The E6 cell, enter the name of “The significance of the correlation coefficients”.
- The cell E7, type designation tobs y, x_1.
- In cell F7 to compute (calculate) from the observations the observed value tobs of the test statistic t for the correlation coefficient of factors y, x_1, enter the formula

= F4*SQRT((20 – 2) : (1 – F4^2)), where 20 – the number of observations, 2 – number of factors.

• In cell E8, enter tobs y, x_2.

• The F8 cell to calculate the correlation coefficient for tobs factors y, x_2, enter the formula

= F5*SQRT((20 – 2) : (1 – F5^2)), where 20 – the number of observations, 2 – number of factors.

• In cell E9 enter tcr.

• In the cell F9 calculate critical tcr follows:

– click on the *fx* button (insert functions);

– in the box *Or select a category*, select *the Statistical*, select a function of the suggested options: select T.INV.2T and then click *OK* (Figure 1). Function *Arguments* window opens. Fill in the fields:

• *X* – enter a value of probability equal 0.05;

• *Deg_freedom* – enter 20–2, where 20 – the number of observations; 2 – number of factors. Click *OK* (Figure 2).

Note – The findings of the existence of dependency and choose the kind of communication functions are described in detail in the section "Econometric analysis of the construction of multiple regression model".

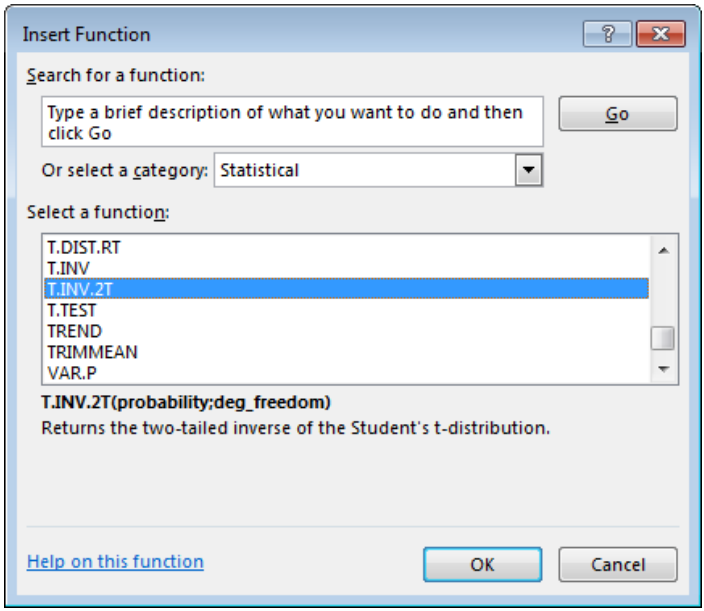


Figure 1 – **Dialog box for insert function**

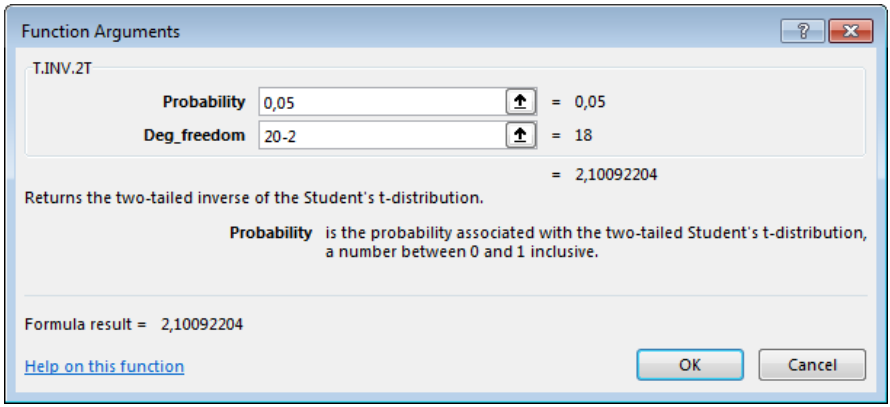


Figure 2 – **Dialog box for function arguments**

3. The parameterization models: finding value estimates pas parameters selected communication functions

Find the least square method estimates of the unknown parameters of the pair of linear regression model $y = b_0 + b_1x_1 + b_2x_2 + \varepsilon$, where ε – random variable, which includes the total effect of all factors unaccounted for in the model by following the steps listed below.

Note – We assume that there exists between the factors for linear-dependence. Next is the linear regression equation. If the proof of nonlinear dependence, the linearization procedure is performed.

Click Data → Data Analysis → Regression → OK. The settings in the dialog box, set the following:

- *Input Y Range* – enter references to cells A1:A21.
- *Input X Range* – enter the cell reference B1:C21.
- *Labels* – check the box.
- *Confidence level* – check the box.
- *Constant is Zero* – leave blank.
- *Output options* – set the switch to a New Worksheet Ply in the appropriate field, enter the name “Regression”.
- *Residuals* – check the box.
- *Standardized residuals* – leave blank.
- *Residual Plots* – check the box.
- *Line Fit Plots* – select the check box.
- *Normal Probability Plots* – leave blank. Click *OK* (Figure 3).

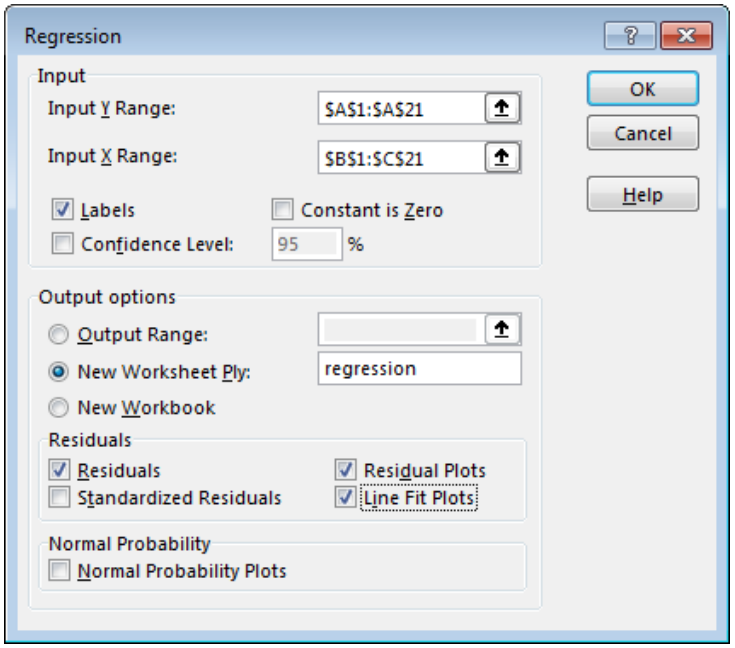


Figure 3 – **Dialog box *Regression***

Place the chart next to (on the graph, click the left mouse button and move the cursor in the white field and the left button pressed move the diagram below) and drag (on the graph, click the left mouse button, the bottom line of the chart while pressing the left button drag down the borders).

Note – The conclusions on the values of the regression equation estimates of the parameters are described in detail in the section "Econometric analysis of constructing a model of the multiple regression".

4. Verification of the model: check the adequacy of the model

4.1. The overall quality of the equation: the importance of checking the coefficient of determination

To test the hypothesis about the importance of coefficient of determination compares the observed value of Fisher statistics found via regression

analysis, and the critical value $F_{a,v_1,v_2} = F_{cr}$, which is calculated on a sheet of "Regression" in the free cell E15 as follows:

- Click on the *fx* button (insert functions).
- In the *Category* box select the *statistical functions* of *the Master*, of the suggested options: F.DIST.RT and then click *OK* (Figure 4).

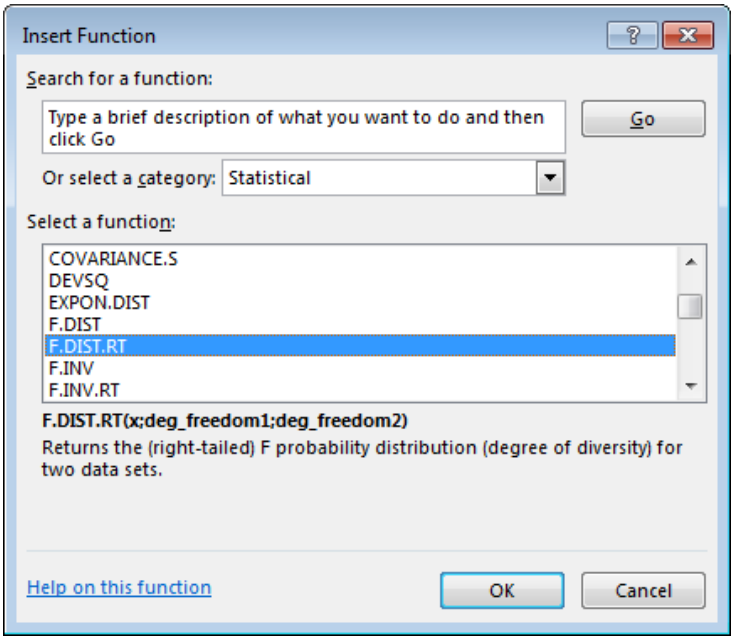


Figure 4 – **Dialog box *Insert function***

Function *Arguments* window opens. Fill in the fields:

- *X* – type in a value of 0.05;
- *Deg_freedom1* – place the cursor in the box and select cell B12 column df of the table ANOVA;
- *Deg_freedom2* – set the cursor and select the cell B13 column df table ANOVA. On-press the *OK* button (Figure 5).

In cell D15 sheet "Regression", enter Fcr.

Note – The conclusions of the equation as described in detail in the section "Econometric analysis of the construction of multiple regression model".

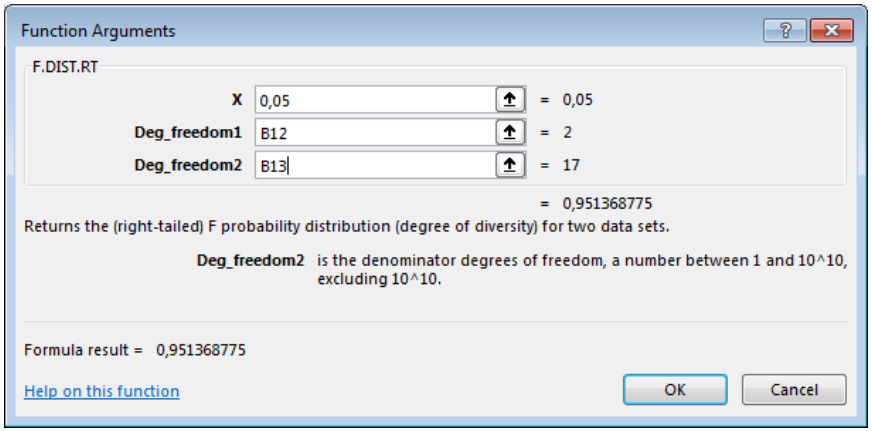


Figure 5 – **Dialog box *Function Argument* for function F.DIST.RT**

4.2. The normality of the distribution of residues: set to allow the use of the Student statistic in hypothesis testing (visual histogram, skewness and kurtosis, using the parametric test hypotheses)

On the sheet "Regression" choose *Data* → *DataAnalysis* → *Descriptive Statistics* → *OK* (Figure 6).

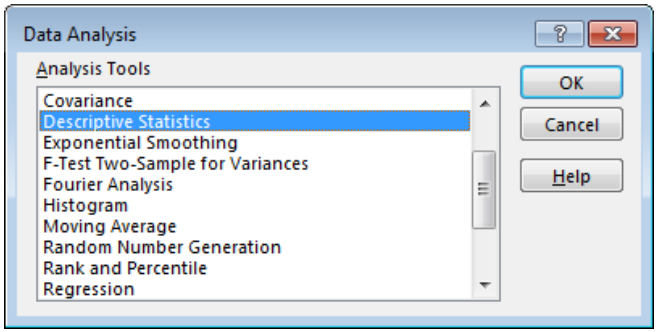


Figure 6 – **Dialog box *Data Analysis***

The parameters set in the dialog box as follows:

- *Input Range* – enter a reference to cell C25:C45 (with the values of the cell residue and the name "Residuals").
- *Grouped by* – check the box columns.
- *Labels in First Row* – check the box.

• *Output Range* – set the switch on the output range in the field next to the mouse pointer to select the cell D26.

• *Summary statistics* – select the check box (Figure 7).

Parameters *Confidence Level for Mean*, *Kth Largest*, *Kth Smallest leave blank*. Click *OK*.

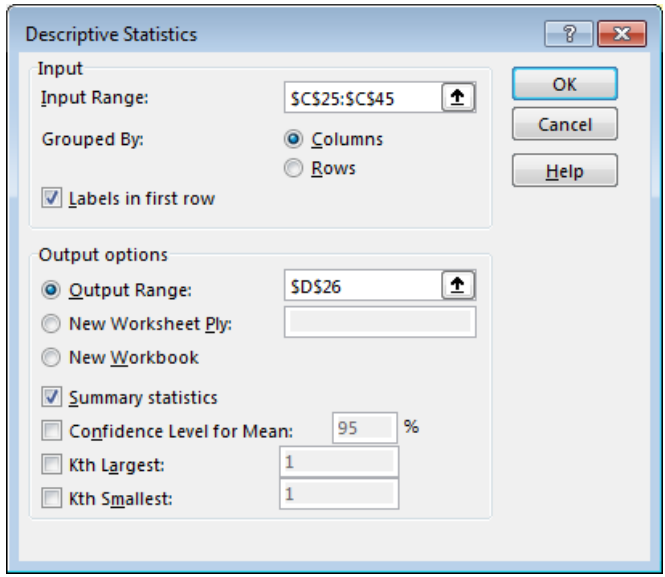


Figure 7 – **Dialog box *Descriptive Statistics***

To test the hypothesis of normal distribution wasps residues using the Pearson's chi-squared test of compared the observed and the critical value of chi-square statistics.

Calculate the observed value of the chi-square statistic $\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$, following the steps below.

In cell A48, enter the name of the "Pearson's chi-squared test".

You take *Data* → *DataAnalysis* → *Histogram* → *OK*. The values of the window, set the parameters as follows:

• *Input Range* – enter a reference to cell C25:C45 (with the values of the cell residue and the title "Residuals").

• *Bin Rang* – do not fill.

• *Labels* – check the box.

• *Output Options* – set the switch Output Range to the output range and enter a reference to cell A49.
• *Pareto* (sorted histogram) – leave blank.
• *Cumulative Percentage* – leave blank.
• *Chart Output* – check the box. Click *OK* (Figure 8).

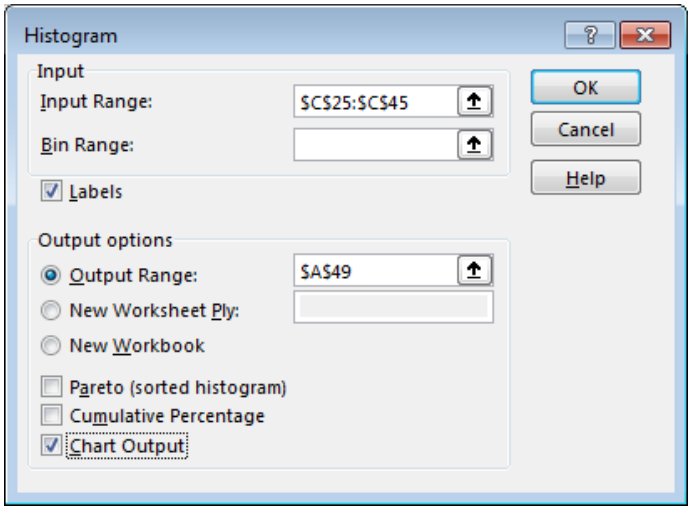


Figure 8 – **Dialog box *Histogram***

Transfer the histogram and drag it down.

Remove the word *More* in the column "Bin" in the same cell, enter the formula = MAX (C26:C45)*3, i. e. the value of the maximum residues have increased three times.

In cell C49, enter the value 0.

In cell C50:C54 enter the array formula:

1) select cell C50:C54;

2) press the key F2;

3) enter the formula = NORM.DIST(A50: A54, E28, E32, TRUE) (Figure 9–10);

4) press the key combination Ctrl + Shift + Enter.

If there was only one value, then press F2 and then Ctrl + Shift + Enter. In the formula bar when activating any cell range C51:C55 will be the formula in curly brackets:

{=NORM.DIST(A50:A54,E28,E32,TRUE)}.

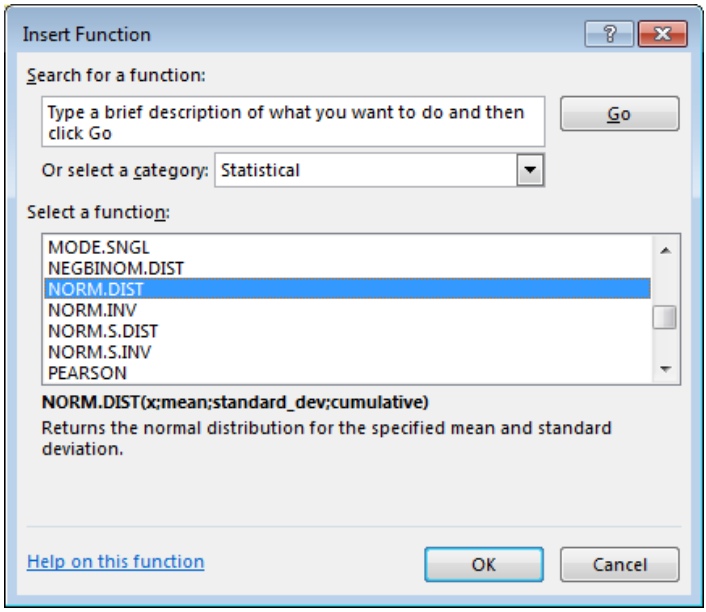


Figure 9 – **Dialog box *Insert Function***

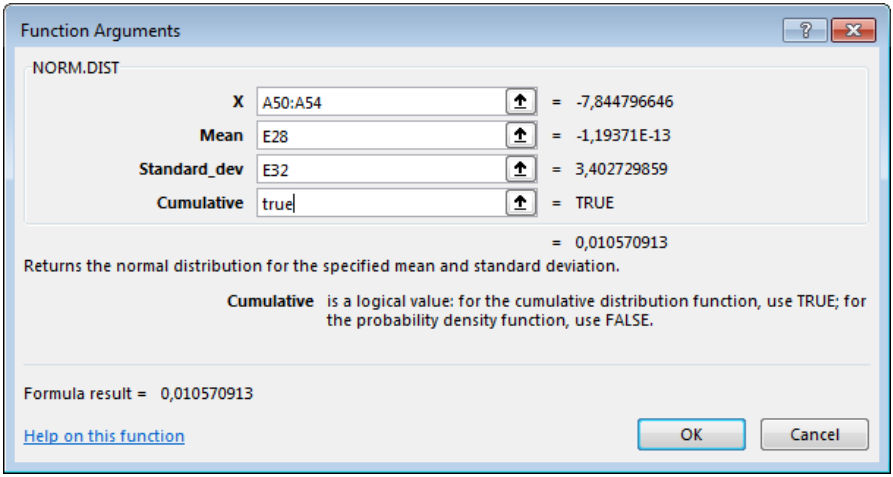


Figure 10 – **Dialog box *Function Arguments* for function NORM.DIST**

Note – In the future, the phrase “enter the array formula” includes a number of four steps: 1) select the range of cells to be filled; 2) press the F2 key on the keyboard; 3) to enter the formula; 4) press the key combination Ctrl + Shift + Enter.

In cells D50:D54, enter the array formula

{=C50:C54–C49:C53}.

In cells E50: E54, type the array formula

{=E40*D50:D54}.

In cell F50: F54 enter the array formula

{=(B50:B54–E50:E54)^2/E50:E54}.

In cell A57, enter the symbol “chi-squared obs”.

In cell B57 enter the formula =SUM(F50:F54) to calculate the chi-square obs.

Find the critical value of the Pearson statistic.

In cell A58, enter the symbol “chi-square cr”.

In cell B58, enter the formula =CHISQ.INV.RT (0.05, 6 – 2 – 1), where 6 – number of intervals; 2 – the number of normal distribution parameters (for calculating the chi-squared test cr) (Figure 11).

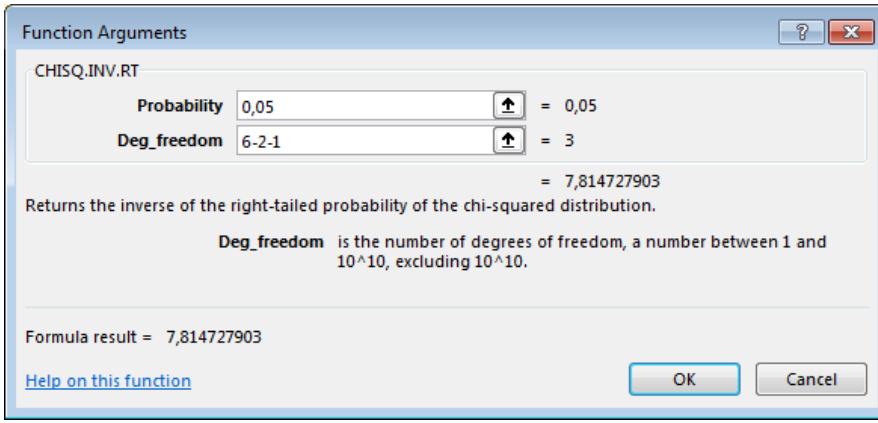


Figure 11 – **Dialog box *Function Argument* for function NORM.DIST**

Note – Conclusions about the normal distribution of residues are described in detail in the section "Econometric Analysis of the Construction of the Multiple Regression Model".

4.3. The significance of the regression coefficients: checking the corresponding hypothesis

To test the significance of the regression coefficients hypothesis compares the observed values of *t*-statistics found via regression analysis, and critical, you want to find as described below.

In cell C20 sheet "Regression", enter tcr. Calculate the critical value tcr free cell D20 as follows:

- Click on the *fx* button (insert functions).
- In the Category box, select the statistical functions of the Master, of the proposed functions below highlight T.INV.2T and then click *OK*.

Function arguments window opens. The parameters are as follows:

– Probability – type in a value of 0.05;

– Degrees of Freedom – enter 20 – 2 – 1, where 20 – the number of observations; 2 – the number of factors in the regression equation; 1 – the number of free membership (b_0) in the regression equation. Click *OK* (Figure 12).

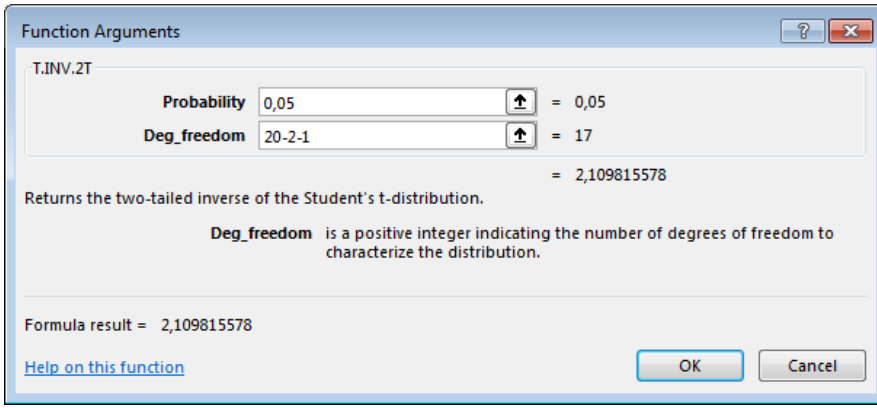


Figure 12 – **Dialog box *Function Argument* for function T.INV.2T**

Note – Conclusions about the significance of regression coefficients are described in detail in the section "Econometric analysis of the Construction of the Multiple Regression Model".

4.4. Checking the statistical properties of the residuals (the quality of regression coefficients)

4.4.1. The centrality of residuals

To test the hypothesis about the importance of the mathematical expectation of a random variable, sample estimates which are residues, compared the observed $t = \frac{(\varepsilon - 0)\sqrt{n}}{S}$ and the critical $t_{a,v} = t_{cr}$ value of *t*-statistics.

On a sheet "Regression" in cell D25, enter the name of the "Condition 1". In cell D41, enter tobs. In cell E41, enter the formula =(E28 – 0)* SQRT(E40)/E32 for calculating the observed value tobs statistics. In cell D42, enter tcr. In cell E42 enter the formula = T.INV.2T(0,05;20 – 2 – 1) to calculate the critical point of the Student distribution tcr.

4.4.2. Homoscedasticity (heteroscedasticity) residuals

To test the hypothesis of homoscedasticity random variable compares the observed and the critical value of *t*-statistics.

Find the observed value by the formula *t*-statistics

$$t = r\sqrt{n-1},$$

where $r = 1 - \frac{6\sum D_i^2}{n(n^2-1)}$ – Spearman's rank correlation coefficient;

D_i – the difference between the rank *xi* and rank *ei* balance module.

On the sheet "Data" contents of cells A1:A21 copy in cell A1 of a new sheet and rename the sheet "Condition2". In cell B1 copy of a sheet of "Regression" column "Residuals" with the name. In cell C1, enter the name "Module of residuals". In cell C2:C21, enter the array formula {=ABS(B2:B21)} (select the cells C2:C21, press the F2 key, enter the formula, press the key combination Ctrl + Shift + Enter).

Select the menu options *Data* → *DataAnalysis* → *Rank and Percentile* → *OK* and fill in the following dialog box as:

- *Input Range* – enter references to cells A1:A21;
- *Labels in first row* – select the check box;
- *Output Range* – D1 cell. Click *OK* (Figure 13–14).

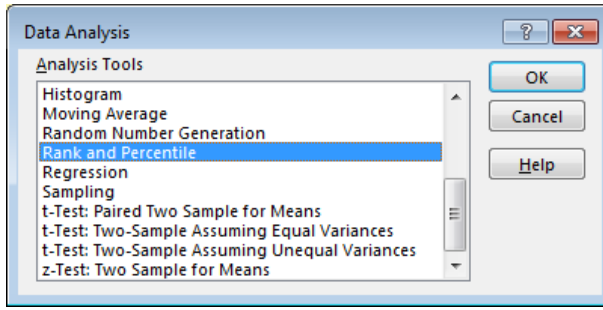


Figure 13 – **Dialog box *Data Analysis***

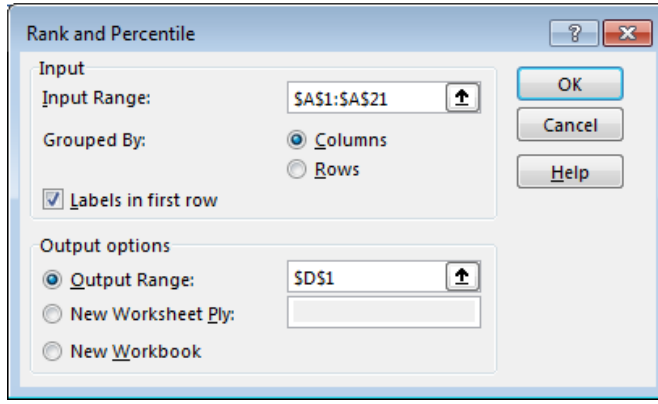


Figure 14 – **Dialog box *Rank and Percentile* for array A1:A21**

Select the menu options *Data* → *Data Analysis* → *Rank and Percentile* → *OK* and fill in the following dialog box as:

- *Input Range* – enter a reference to cell C1:C21;
- *Labels in first row* – select the check box;
- *Output Range* – the H1 cell. Click *OK* (Figure 15).

Select cells D2:G21 and then click *Sort Smallest to Largest* of toolbar. Select the cells H2:K21 and then click *Sort Smallest to Largest*. In cells L2:L21 enter the array formula {= (F2:F21 – J2:J21)^2}. The L1 cell enter the name of the "Ranks squared difference".

The K22 cell enter tobs. The L22 cell enter the formula = (1 – 6*SUM (L2:L21)/(20*(2^20 – 1)))*SQRT(20 – 1) to calculate tobs.

The K23 cell enter tcr. The L23 cell enter the formula =T.INV.2T(0,05;20 – 2) to calculate tcr.

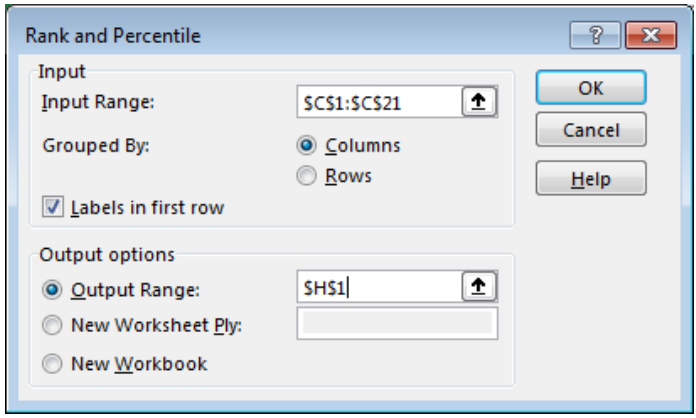


Figure 15 – **Dialog box *Rank and Percentile* for array C1:C21**

4.4.3. Autocorrelation

To test the hypothesis of no autocorrelation slu-tea variable compares the observed and the critical values of the Durbin–Watson statistic.

Find the observed value of statistics $d = \dfrac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}\varepsilon_t^2}$, using as esti-

mates of values of the random variable corresponding values of residues, following the steps listed below.

On the “Regression” sheet in cell F25 enter the name of “Condition3”.

In cell F26:F44 enter the array formula {=(C26:C44 – C27:C45)^2}.

In cell F46 enter the formula = SUM (F26:F44).

In cell G26: G45 enter the array formula {=(C26:C45)^2}.

The G46 cell, enter the formula = SUM (G26:G45).

In cell F47 enter dobs.

The G47 cell enter the formula = F46/G46 to calculate dobs.

If the Durbin–Watson test does not give an answer about the presence of autocorrelation, you can use a visual way to analyze the plot of the residuals from the observation room, built using diagrams (Figure 16).

Note – The conclusions on the implementation of the Gauss-Markov conditions are described in detail in the section “Econometric analysis of the Construction of the Multiple Regression Model”.

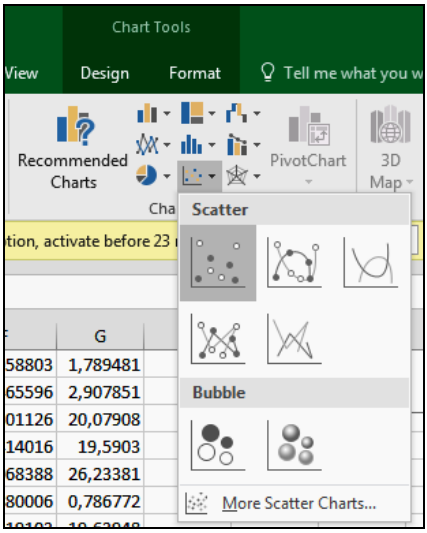


Figure 16 – **Inserting a diagram**

4.5. Analysis of the properties of the model

4.5.1. Multicollinearity factors: identification according explanatory factors

To test the hypothesis of no multicollinearity using the chi-square statistic with $\nu = \frac{n(n-1)}{2}$ degrees of freedom, the observed value of which is determined by the formula

$$\chi^2 = n - 1 - \frac{1}{6}(2p+5)\lg \Delta r,$$

where n – number of observations;

p – the number of independent variables;

Δr – determinant of the matrix of paired coefficients of correlation between factors.

On the sheet “Data” in cell E11, type the name “Multicollinearity”.

In cell E12 type the name of the “Determinant”.

In cell F12 enter the mathematical formula =MDETERM(G4:H5) (Figure 17).

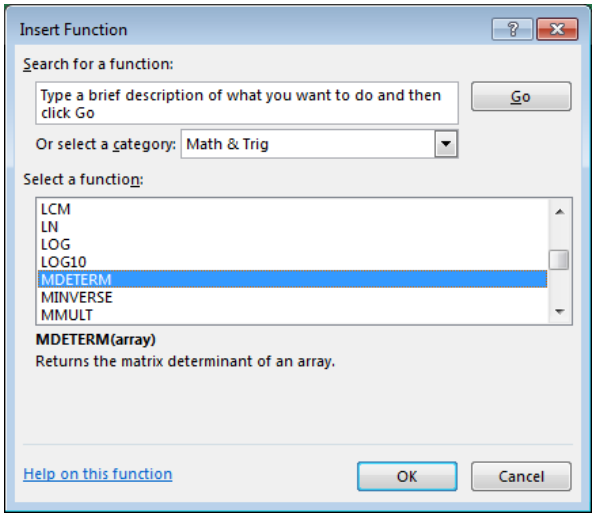


Figure 17 – **Dialog box *Insert Function***

In cell E13, type the name of the "Chi-squared obs".

In cell F13 enter the formula = 20 – 1 – 9*LOG (F12;10)/6 to find the chi-square test, the observed sample.

In cell E14, type the name of the "Chi-squared cr".

In cell F14 enter the formula = CHISQ.INV.RT(0,05; 20*(20 – 1)/2) in order to find the critical chi-square (Figure 18).

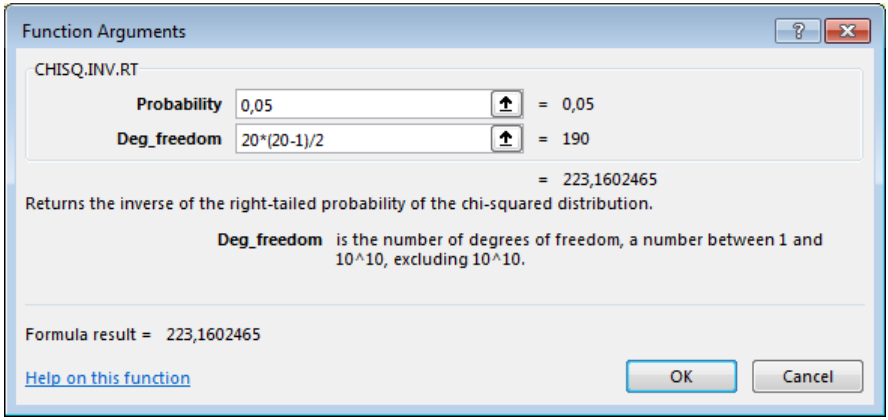


Figure 18 – **Dialog box *Function Argument* for function CHISQ.INV.RT**

4.5.2. Elasticity

The average coefficient of elasticity for *i*-th factor are calculated by linear regression formula.

On the sheet "Data" in cell A23, type the name of the "Elasticity".

In cell A24 enter the name "Mean of y".

In cell A25 enter the formula = AVERAGE (A2:A21) (Figure 19).

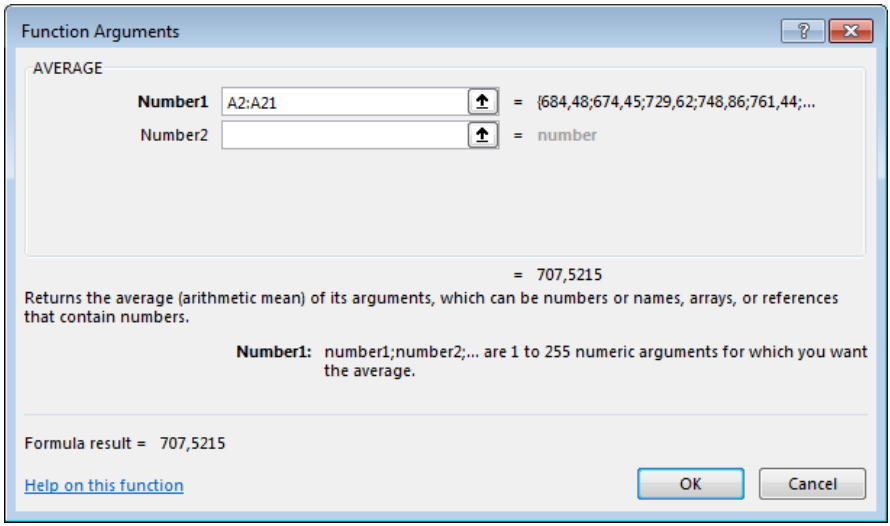


Figure 19 – **Dialog box *Function Argument* for function AVERAGE**

In cell B24, type the name of the "Mean of x1".

In cell B25 enter the formula = AVERAGE (B2:B21).

In cell C24 enter the name "Mean of x2".

In cell C25 enter the formula = AVERAGE (C2:C21).

In cell A26 enter the name of the "Coefficient of x1". From sheet "Regression" copy the coefficient of the variable of factor x1 from cell B18 to cell B26 sheet "Data".

In cell A27 enter the name of the "Elasticity of x1".

In cell B27 enter the formula = B26*B25/A25 for you compute the average private-elasticity of the variable of factor *x*1.

In cell A28 enter the name of the "Coefficient of x2". From sheet "Regression" copy the coefficient of the variable of factor *x*2 from cell B19 to cell B28 sheet "Data".

In cell A29 enter the name of the "Elasticity of x2".

In cell B29 enter the formula = B28*C25/A25 for you compute the average of private-elasticity with respect to factor *x*2.

4.5.3. Private correlation coefficients

Coefficient partial correlation of the first order for the re-variable x1 at a constant value of the variable *x*2 is given by (through steam factor correlation coefficients). To find it, follow these steps:

- in cell E16 type the name of the "Private coefficients of correlation";
- in cell E17 type the name of the "r y, x1-x2";
- in cell F17 enter the formula

$$= (F4 – F5*G5)/SQRT((1 – F5\text{^}2)*(1 – G5\text{^}2)).$$

Similarly find:

- in cell E18 type the name of the "r y, x2-x1";
- in cell F18 enter the formula

$$= (F5 – F4*G5)/SQRT((1 – F4\text{^}2)*(1 – G5\text{^}2)).$$

Testing the significance of partial factors is performed by comparing the observed and the critical values of the *t*-statistic is similar to test the significance of the coefficients of paired correlation at the stage of the specification.

5. Forecasting

The point forecast *y** is found by substituting the values of the explanatory variables 35, 10 into the regression equation.

On a sheet in cell E1 "Regression", enter the name "Point forecast", in cell E2 enter the formula = B17 + B18 + B19*35*10 to calculate the point estimate of the parameter *y* for values of 35 and 10 of the explanatory factors of the conditions of the problem.

Interval prediction or forecast confidence interval is as follows:

$$(y^* - t_{\alpha,\nu}S^*, y^* + t_{\alpha,\nu}S^*),$$

where $t_{\alpha,\nu}$ – the critical value of *t*-statistics for a given level of significance α and the number of degrees of freedom ν;

S^* – the average forecast error standard.

Average standard prediction error calculated by the formula

$$S^* = S \cdot \sqrt{X_p^T (X^T X)^{-1} X_p},$$

where X – the matrix of independent variables observations;

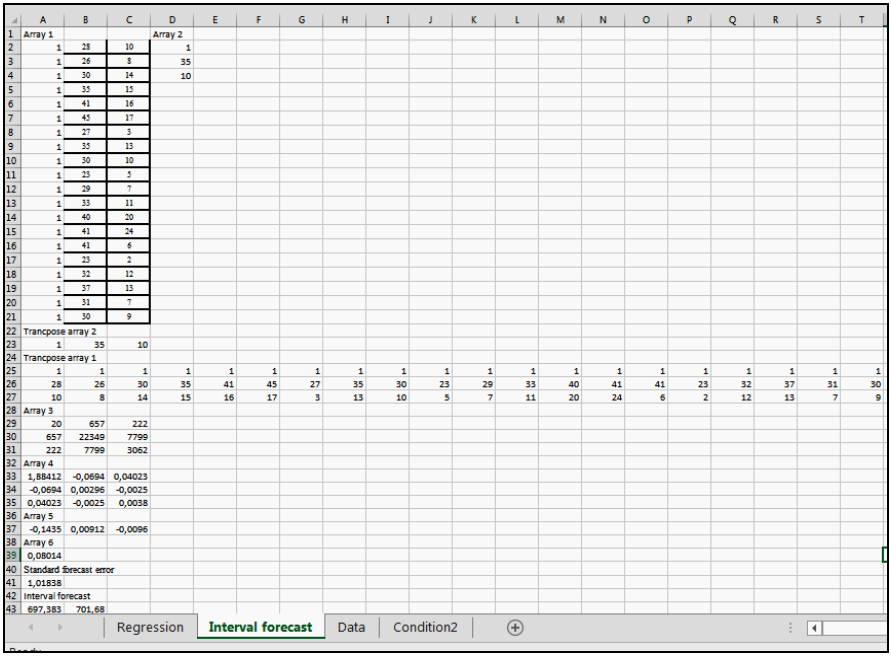
X_p – matrix of values of the independent variables for the pro-prognosis;

S – standard error of the regression;

T – matrix transposition operation.

In cell B2 of the new sheet “Interval forecast” copy cell B2:C21 sheet “Data”. Fill cells A2:A21 units (value of the variable is the free term) (Figure 20). For simplicity further reference in the combined cells A1:C1, enter the name “Array 1” (array X that contains the variable value at the intercept, the factor x1, factor x2, – cells A2:C21), in cell D1 – called “Array 2” (Xp array containing the data to forecast – cells D2:D4). In cell D2 type 1, D3 – 35, D4 – 10.

Example of intermediate calculations standard forecast error and the forecast interval is shown in Figure 20.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Array 1			Array 2																
2	1	28	10	1																
3	1	26	8	35																
4	1	30	14	10																
5	1	35	15																	
6	1	41	16																	
7	1	45	17																	
8	1	27	3																	
9	1	35	13																	
10	1	30	10																	
11	1	23	5																	
12	1	29	7																	
13	1	33	11																	
14	1	40	20																	
15	1	41	24																	
16	1	41	6																	
17	1	23	2																	
18	1	32	12																	
19	1	37	13																	
20	1	31	7																	
21	1	30	9																	
22	Trancpose array 2																			
23	1	35	10																	
24	Trancpose array 1																			
25	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
26	28	26	30	35	41	45	27	35	30	23	29	33	40	41	41	23	32	37	31	30
27	10	8	14	15	16	17	3	13	10	5	7	11	20	24	6	2	12	13	7	9
28	Array 3																			
29	20	657	222																	
30	657	22349	7799																	
31	222	7799	3062																	
32	Array 4																			
33	1,88412	-0,0694	0,04023																	
34	-0,0694	0,00296	-0,0025																	
35	0,04023	-0,0025	0,0038																	
36	Array 5																			
37	-0,1435	0,00912	-0,0096																	
38	Array 6																			
39	0,08014																			
40	Standard forecast error																			
41	1,01838																			
42	Interval forecast																			
43	697,383	701,68																		

Regression | Interval forecast | Data | Condition2

Figure 20 – **Example of calculation of the forecast interval estimation**

To transpose the array, type 2 in cell A23:C23 array formula {=TRANSPOSE(D2:D4)}.

To transpose the array, type 1 in cell A25:T27 array formula {=TRANSPOSE(A2:C21)}.

The result of the product of the transposed array1 dimension 3 on 20 and array 1 dimension 20 on 3 is array 3 dimension 3 on 3, so in cells A29:A31 enter the array formula {=MMULT(A25:T27;A2:C21)}.

The result obtained by calculating the inverse matrix dimension 3 on 3, which is located in cells A33: C35 obtained on formula array {=MINVERS(A29:C31)} (array 4).

The result of the product of the transposed array 2 dimension 1 on 3 and array 4 dimension 3 on 3 is an array 5 of dimension 1 on 3, so in cells A37:C37 enter the array formula {=MMULT(A23:C23;A33:C35)}.

The result of the product array 5 dimension 1 on 3 and array 2 dimension 3 on 1 is an array 6 dimension 1 on 1, so in cell A39 enter the formula =MMULT(A37:C37;D2:D4).

The standard forecast error count in cell A41: =regression!B7* SQRT(A39).

Interval forecast of *y* calculated in cells A43, B43 respectively by the following formulas:

=regression!E2 – regression!D20*'Interval forecast'!A41 (For the left end of the interval);

=regression!E2 + regression!D20*'Interval forecast'!A41 (To the right end of the interval).

Note – Recording 'regression' E2 means that cell E2 is on the list, “regression”. Set and edit the formula in the formula bar is performed.

Econometric analysis of the construction of multiple regression model

In practice the factor *y* depends on many other factors. In the condition of the problem, two most significant factors are identified. There is a task of quantitative description of the dependence of selected economic indicators on the equation of multiple regression on the basis of 20 observations of economic indicators.

The regression type is visually determined by the correlation field, which is depicted on the “regression” sheet.

Since the points are grouped along a straight line (not horizontal), it can be assumed that the dependence of the factor y on the factor *x*1 is linear

and on the factor $x2$ is also linear. It is described by a pairwise linear regression model

$$y = b_0 + b_1 x_1 + b_2 x_2 + \varepsilon,$$

where b_0, b_1, b_2 are unknown model parameters;

ε – a random variable that includes the total influence of all factors not considered in the model.

The coefficient of correlation of the factors y and $x1$ is 0,79 > 0, so the relationship between them is direct and high. The coefficient of correlation of the factors y and $x2$ is 0,99 > 0, so the relationship between them is direct and very high (sheet "Data").

Let's check the importance of the coefficients of pair correlation.

Since | tobs y,x1 | = 5,56 > tcr = 2,1 (sheet "Data"), then the correlation coefficient is significant (significantly different from zero). Consequently, the existence of a linear relationship between the factors y and $x1$ is confirmed.

Since | tobs y,x2 | = 43,79 > tcr = 2,1 (sheet "Data"), then the correlation coefficient is significant. Therefore, the presence of a linear relationship between the factors y and $x2$ is also confirmed.

The point estimate of the parameter b_0 (Y-intersection) is 570,74 (sheet "Regression"), its interval estimate is (560,32; 581,16).

The point estimate of the parameter b_1 for the variable $x1$ is 1,03 (sheet "Regression"), its interval estimate is (0,62; 1,44).

The point estimate of the parameter b_2 for the variable $x2$ is 9,28 (sheet "Regression"), its interval estimate is (8,81; 9,74).

Thus, the regression equation has the following form:

$$y = 570{,}74 + 1{,}03x_1 + 9{,}28x_2.$$

Since any value from the confidence interval can serve as an estimate of the parameter, the regression equation can also have the form:

$$y = 568 + 0{,}8x_1 + 9x_2.$$

Let's estimate the overall quality of the model by the coefficient (index) of determination and the normalized index of determination.

The coefficient of multiple determination R-square is 0,996 (sheet "Regression"). Since it is close to 1, the equation is of high quality. This fact

confirms also the normalized index of multiple determination, equal to 0,996.

Since the observed value of Fobs = 2 392,35 > Fcr = 3,59, then R-square is significant, which again confirms the high quality of the linear multiple regression equation constructed.

Let us analyze the normality of the distribution of residuals for the possibility of using the Student's test when testing statistical hypotheses. The conclusion about the normality of the distribution of residues can be concluded as follows:

- on the histogram of the residues;
- by numerical characteristics of asymmetry and excess;
- according to Pearson's criterion.

Since observed value of chi-squared test 3,16 and less than the critical value of chi-squared test cr 7,81, so the remainders are distributed according to the normal law.

The observed value of the statistics tobs for the coefficient *b*0 is 115,59 (sheet "regression"). The critical value tcr is equal to 2.1. Since | tobs | = 115,59 > tcr = 2,1, then the coefficient b_0 is significant.

Similarly, for the coefficient *b*1, we have the following: tobs = 5,26, tcr = 2.1. Since | tobs | = 5,26 > tcr = 2,1, so the coefficient *b*1 is significant. For the coefficient *b*2 we have: | tobs | = 41,85 > tcr = 2,1, so the coefficient *b*2 is significant.

The importance of regression coefficients confirms the assumption advanced at the stage of the specification of the linear form of the dependence of the factors.

The average residue is –2,84E – 14 = $–2,84 \cdot 10^{–14}$ (sheet "Regression"). It is sufficiently close to zero, so we can assume that condition 1 of Gauss–Markov is satisfied. Let us check the value of the mean for significance, that is, the hypothesis that the mathematical expectation of a random variable is zero.

Let us compare the calculated observed and critical values of the statistics. Since | tobs | = 3,74E – 14 = $3,74 \cdot 10^{–14}$ < tcr = 2,09 (sheet "Regression"), then the average is insignificant (i. e. slightly different from zero). Consequently, the Gauss–Markov condition 1 is satisfied.

Since | tobs | = 1,25 < tcr = 2,1 "Condition 2", then there is no heteroscedasticity. Consequently, condition 2 of Gauss–Markov is satisfied. Hence, estimates of regression parameters will be effective. Therefore, the model can be used for point and interval prediction.

Let us verify the presence of autocorrelation of the residues (Table 2).

Table 2 – **Durbin–Watson statistics: d_1 and d_2 at the level significance of 5% (fragment of the table)**

n	$k = 1$		$k = 2$		$k = 3$		$k = 4$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83

According to the table of critical values (Table 2) of *d*-statistics for the number of observations 20, the number of explanatory variables 2 and the specified significance level 0,05, the values $d_1 = 1{,}10$ and $d_2 = 1{,}54$, which break the interval [0; 4] into five areas (Figure 21).

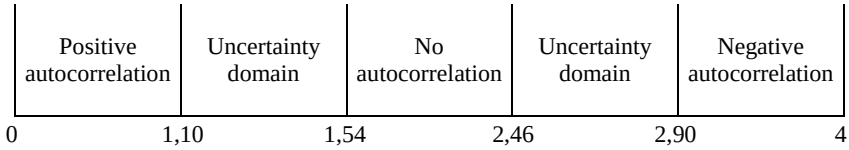


Figure 21 – **Critical regions of Durbin–Watson statistics**

Since 1.10 < dobs = 1,38 < 1,54 (sheet "regression"), i. e. the observed value falls into the zone of uncertainty, nothing can be said about the presence of autocorrelation using the Durbin–Watson criterion.

Visually, the presence of autocorrelation of residues can be determined from the graph of residues obtained on the sheet "Regression".

Since the points on the remainder graph are scattered along the straight line $y = 0$ is chaotic with no apparent regularity, then no relationship between the residues is observed. Therefore, condition 3 is satisfied.

Since the pair correlation coefficient of factors $x1$ and $x2$ r_{x1}, $x_2 = 0{,}75 < 0{,}8$ (sheet "Data"), the relationship between the factors exists, but it is insignificant.

Since the chi-square observed is 19,53 and less than the chi-square critical, equal to 223,16, the multicollinear factors are absent.

With the change in the value of factor $x1$ by 1% for a fixed value of factor $x2$, the value of factor y increases by 0,05% (elasticity of the factors $x1$ and $x2$ on sheet "Data"). Similarly, with a change in the value of factor $x2$ by 1% for a fixed value of factor $x1$, the value of factor y increases by 0,15%. Hence, the influence of the factor $x2$ is greater than the factor $x1$.

Since partial correlation coefficients 0,78 < 0,995, then from two factors are more influenced the factor *x*2.

Both partial correlation coefficients are significant: | tobs (r y, x1–x2) | = 5,4 > tcr = 2,1, | tobs (r y, x2 – x1) | = 43,06 > tcr = 2.1.

Conclusion based on the results of the verification phase: since all verification conditions are fulfilled, the model is qualitative. Thus, the forecast made on it is qualitative, that is, unbiased, consistent and effective.

The point forecast of the factor y is 699,53. The interval forecast (697,38; 701,68) (sheet "Interval forecast") means that with a probability of 0,95 any value from this interval is an estimate of the factor *y*.

Questions for self-control

1. What are the stages of building an econometric model?
2. What is the brief characteristic of the goal of each stage?
3. Knowledge of what scientific disciplines are needed at each stage of econometric modeling?
4. What is the specification of the multiple regression model?
5. Why is there a random variable in the regression equation?
6. How to determine the strength and direction of the interaction of factors?
7. What does the correlation coefficient mean?
8. How to check the correlation coefficient for significance?
9. What is the essence of LQM for finding estimates of regression parameters?
10. Why are LQM estimates of the parameters, and not their exact values?
11. What is the point estimate for a parameter?
12. What is the essence of interval estimation of parameters?
13. How to find interval estimates of regression coefficients?
14. How are standard regression errors and standard errors of regression coefficients used when analyzing estimates of regression parameters?
15. What is the economic meaning of the parameters of the regression model?
16. How can we evaluate the overall quality of the regression equation?
17. What is the essence of the coefficient of determination, the normalized coefficient of determination? To what extent do they change?
18. What is the relationship between the coefficient of determination, the correlation coefficient and the multiple correlation coefficient for multiple regression?

19. For what purpose is the Fisher test used in the pair regression?

20. How to get the remainder for the pair regression model?

21. What condition should the residuals satisfy, so that Student's criterion can be used to test statistical hypotheses?

Individual task

Proceed as follows:

- investigate the dependence of factor *y* on factors *x*1 and *x*2, using the observations given in Table 1, adding to the values of factor *y* a value of 10k, where *k* is a student personal number in the students journal;
- build a regression model;
- calculate the value of the factor *y* if *x*1 = 35 and *x*2 = 10;
- make a report.

THEME 2. NONLINEAR REGRESSION

Nonlinear regression is divided into the following two types:

- a regression what is nonlinear for the explanatory variable, but linear for the estimated parameters (eg, polynomial $\tilde{y} = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$, hyperbole $\tilde{y} = a + \frac{b}{x}$);
- a regression what is nonlinear for the evaluated parameters (for example, a power model $\tilde{y} = a \cdot x^b$, an exponential model $\tilde{y} = a \cdot b^x$).

To determine estimates of parameters of paired nonlinear regression models is used linearization procedure. It consists by transformations of variables of the equations and then to analyze the obtain linear equation with new variables.

"Nonlinear" explanatory variables are replaced by new "linear" variables. After that, to a new regression ordinary least square method is applied.

To estimate the regression parameters of nonlinear on estimated parameters, commonly used method is the logarithm with subsequent replacement of variables.

Table 3 shows the types of the regression and formulas to estimate parameters of regression models.

Table 3 – **Estimates of the parameters of nonlinear regression models**

Type of regression formula	Linearizing transformation	Parameters of regression model
Exponential regression $\tilde{y} = e^{ax+b}$	$x' = x;$ $y' = \ln y$	$b = \dfrac{n\sum_{i=1}^{n}(x_i \ln y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \ln y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$ $a = \dfrac{1}{n}\sum_{i=1}^{n} \ln y_i - \dfrac{1}{n} b \sum_{i=1}^{n} x_i$
Logarithmic regression $\tilde{y} = a + b \ln x$	$x' = \ln x;$ $y' = y$	$b = \dfrac{n\sum_{i=1}^{n}(\ln x_i \cdot y_i) - \sum_{i=1}^{n} \ln x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} (\ln x_i)^2 - \left(\sum_{i=1}^{n} \ln x_i\right)^2};$ $a = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{1}{n} b \sum_{i=1}^{n} \ln x_i$
Power regression $\tilde{y} = a \cdot x^b$	$x' = \ln x;$ $y' = \ln y$	$b = \dfrac{n\sum_{i=1}^{n}(\ln x_i \cdot \ln y_i) - \sum_{i=1}^{n} \ln x_i \sum_{i=1}^{n} \ln y_i}{n\sum_{i=1}^{n} (\ln x_i)^2 - \left(\sum_{i=1}^{n} \ln x_i\right)^2};$ $\ln a = \dfrac{1}{n}\sum_{i=1}^{n} \ln y_i - \dfrac{1}{n} b \sum_{i=1}^{n} \ln x_i$
Indicative regression $\tilde{y} = a \cdot b^x$	$x' = x;$ $y' = \ln y$	$\ln b = \dfrac{n\sum_{i=1}^{n}(x_i \ln y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \ln y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$ $\ln a = \dfrac{1}{n}\sum_{i=1}^{n} \ln y_i - \dfrac{1}{n} \ln b \sum_{i=1}^{n} x_i$
Hyperbolic regression $\tilde{y} = a + \dfrac{b}{x}$	$x' = \dfrac{1}{x};$ $y' = y$	$b = \dfrac{n\sum_{i=1}^{n}\dfrac{y_i}{x_i} - \sum_{i=1}^{n}\dfrac{1}{x_i}\sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n}\dfrac{1}{x_i^2} - \left(\sum_{i=1}^{n}\dfrac{1}{x_i}\right)^2};$ $a = \dfrac{1}{n}\sum_{i=1}^{n} y_i - \dfrac{1}{n} b \sum_{i=1}^{n}\dfrac{1}{x_i}$

Formulation of the problem

Examine the relationship between factors *y* and *x*, using observations presented in Table 4. Determine a regression model $y = f(x) + \varepsilon$. Calculate the value of the index *y* when *x* = 230.

Table 4 – **Observations**

Factor	Values										
1	2	3	4	5	6	7	8	9	10	11	12
x	110	125	132	137	160	177	192	215	235	240	245
y	63	85	126	165	284	678	752	1 310	2 697	3 137	3 641

Continued table 4

Factor	Values					
1	13	14	15	16	17	18
x	250	275	285	295	320	344
y	5 230	9 555	12 190	21 318	44 644	86 569

Computing technology in MS Excel to building nonlinear regression model and its econometrics analysis

Prepare data for calculations (enter given data presented in Table 4). In the cell A1 enter the title *Factor x* and in the cell B1 enter *Factor y*. In the cells A2, …, A18 enter the values of factor *x*, in the cells B2, …, B18 enter the appropriate values of factor *y*.

The type and strength of the functional dependence (linear or nonlinear) are determined by covariance and correlation coefficient.

On the *Data tab* click tab item *Data Analysis* and then select *Correlation*. Set the parameters as follows:

- *Input Range* – given data in the cells A1:B18.
- *Grouped By* – columns.
- *Labels in first column* – check the box.
- *Output options* – check the box Output Range and in the field next select the cell A20. Click on *OK*.

To test the hypothesis on the significance of the correlation coefficient the observed value $t = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}}$ and the critical value $t_{\alpha,\nu} = t_{cr}$ of the Student's statistics are compared. To calculate them do the following:

- In the cell E21 enter the title tobs.
- In the cell F21 enter the formula

$$=B22*SQRT(17 – 2)/SQRT(1 – B22^2)$$

to calculate the observed value tobs, where 17 is the number of observations; 2 – the number of factors.

- In the cell E22 enter tcr.
- In the F22 calculate the critical value tcr using the function T.INV.2T with probability $\alpha = 0{,}05$ and Deg_freedom equals to 17 – 2, where 17 – the number of observations; 2 – the number of factors.

As a result of calculations a correlation matrix is obtained (Table 5).

Table 5 – **Correlation matrix**

Correlation		
	Factor x	*Factor y*
Factor *x*	1	
Factor *y*	0,71	1

The correlation coefficient between the factors is equal to 0,71 > 0, so the dependence between them is direct and high by Chaddock scale.

We will verify the significance of the correlation coefficient, as it is found only on 17 observations, so it may lead to wrong conclusions about the entire population factors. The observed value and the critical value of the Student's statistics are obtained (Table 6).

Table 6 – **Significance of the correlation coefficient**

tobs	3,94
tcr	2,13

As | tobs | = 3,94 > tcr = 2,13 then the correlation coefficient is significant (significantly different from zero).

Based on the analysis, we put forward the hipothesis that *y* dependence of *x* is described by a linear regression model $y = b_0 + b_1 x + \varepsilon$, where b_0, b_1 – the unknown parameters of the model, ε – a random variable, which includes the total effect of all unaccounted factors in the model and measurement error.

We need to find the least squares estimates of the parameters b_0, b_1 in the regression model. On the *Data tab* click tab item *Data Analysis* and then select *Regression*. In the dialog box set the parameters as follows:

- *Input Y Range* – select the cells B1:B18.
- *Input X Range* – select the cells A1:A18.
- *Labels* – check the box.
- *Confidence level* – check the box.
- *Constant is Zero* – clear the check box.
- In the *Output option*s select *Output Rang* and enter cell reference A24.
- *Residuals* – check the box.
- *Standardized Residuals* – clear the check box.
- *Residuals Plots* – check the box.
- *Line Fit Plots*– check the box.
- *Normal probability Plots* – clear the check box.

Click on *OK*. The scatter plot and the fitted trendline are on Figure 22. Each column in the data table (Table 4) is represented by a marker whose position depends on its values in the columns set on the *X* and *Y* axes.

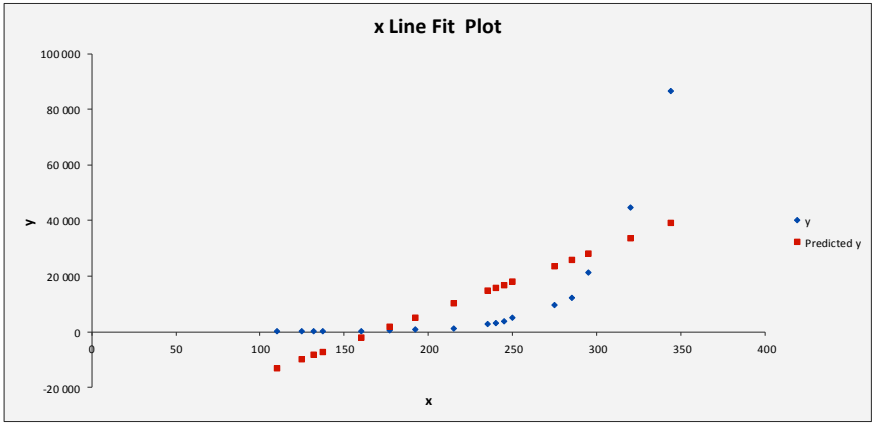


Figure 22 – **Scatter plot**

As the points of the scatter plot are far from the trendline, we can conclude that the relationship between the factors *x* and *y* are not described by paired linear regression $y = b_0 + b_1 x + \varepsilon$.

According to the position of points on the scatter diagram it can be assumed that the relationship can be described by an exponential model $y = e^{ax+b} + \varepsilon$, where *a*, *b* are unknown parameters, ε – an error term.

Despite the obtained contradiction of regression analysis point and interval estimation on individual regression coefficients are given in the third table of output of results.

Table 7 – **Statistics of the regression coefficients**

	Coefficients	Standard Error	*t*-Stat	*P*-value	Lower 95%	Upper 95%
Intercept	–37 722,55	13 053,1	–2,89	0,01	–65 544,6	–9 900,53
x	223,10	56,6	3,94	0,001	102,4	343,76

The point estimation on the parameter b_0 is equal to –37 722,55. The point estimation on the parameter b_1 is equal to 223,10. Thus the simple linear regression model is written as

$$y = –37\,722{,}55 + 223{,}10x.$$

It is necessary to check the quality of the obtained linear regression model.

To check the significance of R-squared statistic of the regression is to use an F test. The *F*-statistic calculated from the data is compared with the critical value $F_{\alpha,\nu_1,\nu_2} = F_{cr}$, of the *F*-distribution. The *F*-statistic is given by the *Data analysis*. Its critical value is calculated in the cell E38 by the function =F.INV.RT(0,05;B35;B36), where *Probability* is equals to 0,05, *Deg_freedom 1* is equal to 1 (this value is in the cell B35) and *Deg_freedom 2* is equal to 15 (this value is in the cell B36). In the cell D38 enter the label Fcr.

In the table “Regression Statistics” there are the indicators describing the general quality of the regression model (Table 8).

For the simple regression model R-squared statistic is equal to the square of the correlation coefficient between the factors and is equals to 0,51. The coefficient of multiple determination is not close enough to 1, so the model adequacy is not good (the quality of the regression model is sa-

tisfactory if R-squared is greater than 0,7). To determine more precisely the quality of the model allows the adjusted (normalized) R-squared. Adjusted R-squared is equal to 0,48 which confirms the poor quality of the model and refutes the assumption that there is a linear relationship between the factors.

Table 8 – **Regression statistics**

SUMMARY OUTPUT	
Regression Statistics	
Multiple R	0,71
R-Square	0,51
Adjusted R-Square	0,48
Standard Error	16 251,21
Observations	17

Significance of R-squared statistic of the regression model is determine with an F test which is calculated in the table "Variance analysis" (Table 9).

Table 9 – **Variance analysis**

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4 102 191 643	4 102 191 643	15,53261	0,001306885
Residual	15	3 961 527 340	264 101 823		
Total	16	8 063 718 983			
			Fcr	4,54	

As the *F*-test calculated 15,53 is greater that the critical value of the *F*-distribution Fcr = 4,54, R-squared statistic is statistically significant (significantly different from zero), but it not sufficiently close to 1 that confirms doubt about regression model.

As there is the relationship between factors, we will check the hypothesis that the relationship can be described by an exponential function

$$y = e^{ax+b} + \varepsilon,$$

where *a*, *b* are unknown parameters of the model.

To determine estimates of the parameters of the exponential model of pair regression the procedure of linearization is used. To evaluate the parameters of exponential regression, nonlinear in the estimated parameters, the logarithmic method is used, with the subsequent replacement of the variable y by the variable ln y.

Convert the values of the variable y. To do this, follow these steps:

- in cell C1 enter the name of the new variable z = ln y;
- in cell C2, enter the formula =ln(B2) and use the autocomplete, copy this formula into cells C3:C18.

Verify, using the correlation coefficient and the correlation field, that the relationship between the variables x and z is linear. Regarding the new set of variables x and z, construct a pair linear regression model, verify it, find the predicted value of the factor y for the value of the explanatory factor given in the condition.

To obtain the predicted value of the factor y, it is first necessary to obtain the value of z for a given value of the factor $x = 23$, and then to find $y = e^z$.

Questions for self-control

1. What are the types of nonlinear regression models?
2. What are the examples of regression which are nonlinear for the explanatory variable, but linear for the estimated parameters?
3. What are the examples of regression which are nonlinear in the evaluated parameters?
4. What method is used to estimate the regression parameters of nonlinear equation for the inclusion in the explanatory variable, but linear in the estimated parameters?
5. What is the principle of the method change of variables in the linearization of nonlinear regression?

Individual task

Examine the relationship between factors y and x, using observations presented in Table 10 where k is a student personal number in the students journal. Determine a regression model $y = f(x) + \varepsilon$. Calculate the value of the index y when $x = 230 + 10 \cdot k$.

Table 10 – **Given data**

Factor x	Factor y
$110 + 10 \cdot k$	63
$125 + 10 \cdot k$	85
$132 + 10 \cdot k$	126
$137 + 10 \cdot k$	165
$160 + 10 \cdot k$	284
$177 + 10 \cdot k$	678
$192 + 10 \cdot k$	752
$215 + 10 \cdot k$	1 310
$235 + 10 \cdot k$	2 697
$240 + 10 \cdot k$	3 137
$245 + 10 \cdot k$	3 641
$250 + 10 \cdot k$	5 230
$275 + 10 \cdot k$	9 555
$285 + 10 \cdot k$	12 190
$295 + 10 \cdot k$	21 318
$320 + 10 \cdot k$	44 644
$344 + 10 \cdot k$	86 569

THEME 3. TIME SERIES

1. Problem definition. Data collection

A *time series* is a time-oriented or chronological sequence of observations on a variable of interest. Time series data are indexed by time. Typical examples include macroeconomic aggregates, prices and interest rates. This type of data is characterized by serial dependence so the random sampling assumption is inappropriate. Most aggregate economic data is only available at a low frequency (annual, quarterly or perhaps monthly). The exception is financial data where data are available at a high frequency (weekly, daily, hourly, or by transaction) so sample sizes can be quite large.

A time series y_t is a process observed in sequence over time, $t = 1, \ldots, T$. To indicate the dependence on time, we use the subscript t to denote the individual observation and T to denote the number of observations.

Quantitative forecasting techniques make formal use of historical data and a forecasting model. The forecasting model is used to extrapolate past and current behavior into the future.

2. Specification: model selection

2.1. Determination of basic feature of the time series

Time series plots can reveal patterns such as random, trends, levels shifts, periods or cycles, unusual observations, or a combination of patterns. The general mathematical model for decomposition of a time series into components is

$$y_t = f(T_t, S_t, C_t, \varepsilon_t),$$

where T_t – the trend effect;
S_t – the seasonal component;
C_t – the cyclic component;
ε_t – the random error component.

There are usually two forms for the function f: an additional model

$$y_t = T_t + S_t + C_t + \varepsilon_t$$

and a multiplicative model

$$y_t = T_t \cdot S_t \cdot C_t + \varepsilon_t.$$

The additive model is appropriate if the magnitude (amplitude) of the seasonal variation does not vary with the level of the series, while the multiplicative version is more appropriate if the amplitude of the seasonal fluctuations increases or decreases with the average level of the time series.

The model according the original data contains at least one component.

2.1.1. Graphical display

The basic graphical display for time series data is the time series plot. This is just a graph of y_t versus the time period, for $t = 1, 2, \ldots, T$. Features such as trend and seasonally are usually easy to see from the time series plot.

2.1.2. Determination of basic feature of the time series using autocorrelation function and correlogram

If a time series is stationary this means that the join probability distribution of any two observations, say, y_t and $y_{t+\tau}$, is the same for any two time

periods t and $t + \tau$ that are separated by the same interval τ. The interval τ is called the *lag*.

The autocorrelation coefficient at lag τ for the stationary time series is

$$r_\tau = \frac{\sum_{t=1}^{T-\tau} (y_t - \bar{y}_1)(y_{t+\tau} - \bar{y}_2)}{\sqrt{\sum_{t=1}^{T-\tau} (y_t - \bar{y}_1)^2 \cdot \sum_{t=\tau+1}^{T} (y_{t+\tau} - \bar{y}_2)^2}},$$

where $\bar{y}_1 = \frac{\sum_{t=1}^{T-\tau} y_t}{T-\tau}$, $\bar{y}_2 = \frac{\sum_{t=\tau+1}^{T} y_t}{T-\tau}$.

The maximum lag should not be less than n/4 to ensure the statistical validity of autocorrelation coefficients.

Autocorrelation coefficient is similar to the linear correlation coefficient and varies between –1 and 1. Its significance is assessed using t-statistic $t = r_\tau \sqrt{\frac{n-2}{1-r_\tau^2}}$, which has Student distribution with $k = n-2$ degrees of freedom. If $|t| \geq t_{\alpha,k}$, the coefficient is significant.

The time series structure is analyzed on the basis of the autocorrelation coefficients. If first-order autocorrelation coefficient is maximal the time series includes only trend and random component. If the highest autocorrelation coefficient is of order t, the time series contains the cyclical fluctuations with frequency t. If none of the autocorrelation coefficients is not significant, neither the time series does not contain trend but only cyclic pattern, or the time series contains strong nonlinear trend.

The collection of values of r_τ is called the *autocorrelation function* (*ACF*). Graphical display of ACF is called *correlogram*.

2.2. Determination of model type: additional model or multiplicative model

If there are several components in the time series, it is necessary to determine how to combine them into a model: to multiply or to add. The time series plot is analyzed. If the amplitude of the oscillation decreases, the

time series is described by a multiplicative model, if the oscillation amplitude is constant, an additive model is used.

3. Analytical alignment of time series

3.1. The structural stability of the time series

There are one-time changes in the nature of the time series caused by structural changes in the economy or by other factors. They differ from the seasonal and cyclical fluctuations because from a certain moment of time the nature of the dynamics of the studied index changes. It is necessary to change trend parameters describing the dynamics (Figure 23). In this case piecewise-linear model is used.

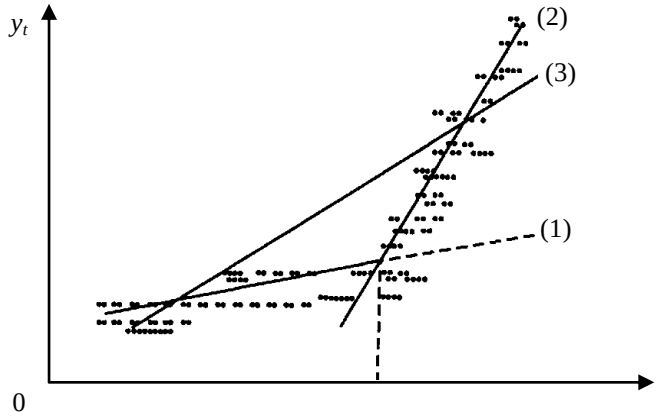


Figure 23 – **The nature of the trend of the time series has changed at moment t^***

3.2. Time-series alignment

One of the most common ways of time series modeling is to determine trend or analytical function characterizing the time series. This analytical method is called time-series alignment. The following functions of time can be used for analytic alignment:

- linear function $\tilde{y}_t = a + b \cdot t$;
- hyperbolic function $\tilde{y}_t = a + \frac{b}{t}$;
- exponential function $\tilde{y}_t = e^{a+b \cdot t}$;
- power function $\tilde{y}_t = a \cdot t^b$;

• high order polynomial $\tilde{y}_t = a + b_1 \cdot t + b_2 \cdot t^2 + ... + b_k \cdot t^k$.

The models are usually fit to the data by using ordinary least squares.

4. Verification

To choose the best model is to verify basic functions of time, to calculate for each function R^2 statistic and the Fisher test and then to choose the model that maximizes the R^2 statistic. The implementation of this method is relatively simple when the computer software is used.

Formulation of the problem

Table 11 shows annual productions of a product.

Table 11 – **Original observations, rubles**

Year	Annual production	Year	Annual production
1990	5 665	2000	19 037
1991	9 570	2001	21 748
1992	11 172	2002	23 298
1993	10 150	2003	26 570
1994	12 704	2004	26 080
1995	12 588	2005	27 446
1996	13 018	2006	29 658
1997	13 471	2007	32 573
1998	15 017	2008	36 435
1999	17 356	2009	38 100

Analyze the structure of the time series, check out the hypothesis of its structural stability, execute analytical smoothing time series, and make a forecast for 2011.

Computing technology in MS Excel to building time series and its econometrics analysis

1. Problem definition. Data collection

In cell A1 enter the name of the first column “Year”, in cell B1 the name of the second column “Annual production”. In the cells A2, A3, ...,

A21 enter the first column data of the Table 11, using auto-complete feature, in cell B2, B3, ..., B21 enter second column data.

Rename the sheet Given data. Save the book under the title Time series.

Annual productions of a product shown in Figure 22 are a time series. The task of the quantitative forecasting technique is to specify a formal model where *t* is an argument and *y* is a dependant variable.

2. Specification: model selection

2.1. Determination of basic feature of the time series

2.1.1. Graphical display

Construct the time series plot to show specific feature of a time series. The time series plot is just a graph of observations given in Table 11.

Figure 24 shows time series plot for annual productions. As we can see the time series contains increasing trend.

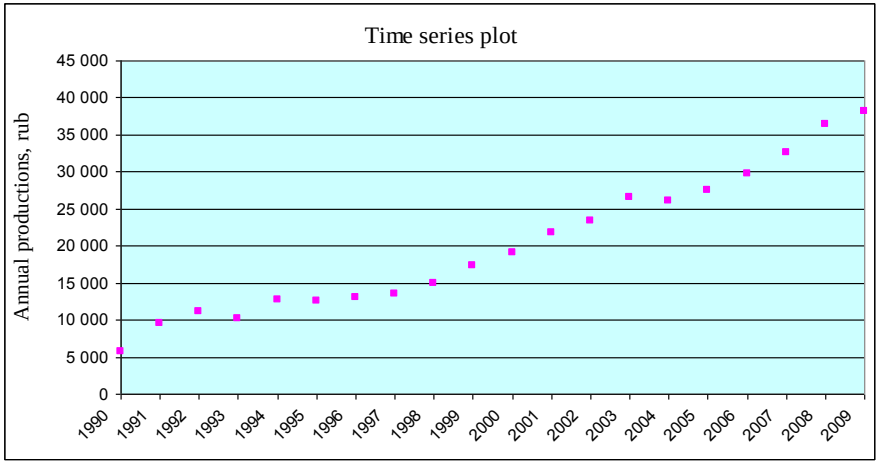


Figure 24 – **Time series plot for annual productions**

2.1.2. Determination of basic feature of the time series using autocorrelation function and correlogram

To ensure statistical significance of autocorrelation coefficients, determine the number of lags to *s*. The maximum lag should not be less than n/4 = 20/4 = 5.

On the sheet “Given data” enter “Lag” in the cell D1. In the cells D2:D6 enter 1, 2, …, 5 using auto-complete feature.

In the cell E1 enter “Autocorrelation coefficient”.

In the cell E2 enter the formula below to calculate autocorrelation coefficient r1

=CORREL(B2:B20,B3:B21).

In the cell E3 enter the formula below to calculate autocorrelation coefficient r2

=CORREL(B2:B19,B4:B21).

In the cell E4 enter the formula below to calculate autocorrelation coefficient r3

=CORREL(B2:B18,B5:B21).

In the cell E5 enter the formula below to calculate autocorrelation coefficient r4

=CORREL(B2:B17,B6:B21).

In the cell E6 enter the formula below to calculate autocorrelation coefficient r5

=CORREL(B2:B16,B7:B21).

To evaluate the significance of autocorrelation coefficients do the following:

- in the cell F1 enter “Sample size”;
- in the cells F2:F6 enter 19, 18, …, 15 using auto-complete feature;
- merge cells G1 and H1 and enter *Significance of autocorrelation coefficients* in the merged cell;
- in the cells G2:G6 enter t1, t2, …, t5 respectively;
- in the cell H2 enter the formula =E2*SQRT((F2 – 2)/(1 – E2^2));
- in the cells H3: H6 fill in by similar formulas using auto-complete feature;
- in the cell D7 enter MAX;
- in the cell E7 enter the formula =MAX(E2:E6).

Calculate the critical value as follows:

- in the cell I1 enter tcr;
- in the cell I2 enter the formula =T.INV.2T(0.05, F2-2).

Cells I3:I6 fill in by similar formulas using auto-complete feature.

If the biggest is the autocorrelation coefficient at lag $\tau > 1$, the time series has patterns such as random, trend and seasonal variation; if $\tau = 1$ the time series has only trend and random variation.

Construct the correlogram (Figure 25).

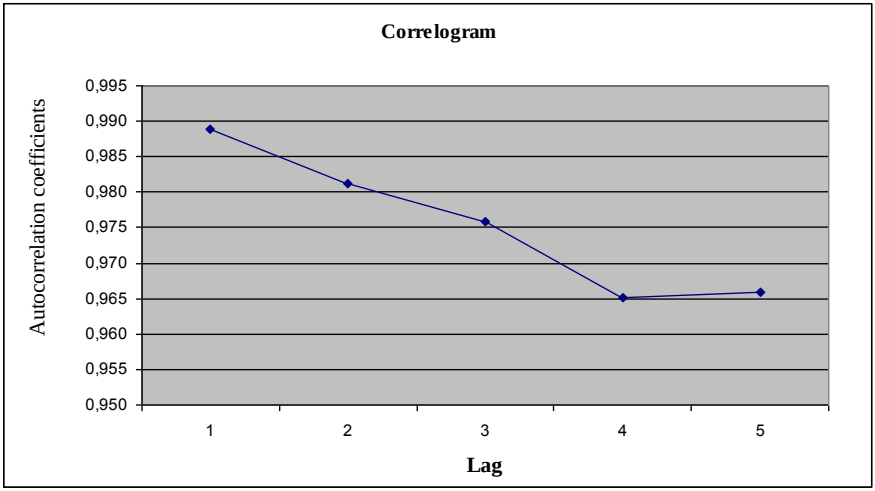


Figure 25 – **Correlogram**

In our example first-order autocorrelation coefficient is maximal, so the time series includes only trend and random component and does not contain cyclic and seasonal patterns (Table 12 and Figure 25).

Table 12 – **Values of ACF**

Lag (τ)	Autocorrelation coefficients (r_τ)
1	0,989
2	0,981
3	0,976
4	0,965
5	0,966
Max	0,989

As all values of the *t*-statistic are above the critical values (Table 13), all the autocorrelation coefficients are significant. This confirms the hypothesis that the time series contains trend and random components.

Table 13 – **Significance of autocorrelation coefficients**

Sample size	Autocorrelation coefficients significance		t_{cr}
19	t_1	27,360	2,110
18	t_2	20,338	2,120
17	t_3	17,316	2,131
16	t_4	13,813	2,145
15	t_5	13,431	2,160

2.2. Determination of model type: additional model or multiplicative model

As we can see on Figure 24 the time series has no fluctuations. Consequently, its model is additive and represents the sum of the trend and the random component.

3. Analytical alignment of time series

3.1. The structural stability of the time series

As Figure 24 shows the nature of the dynamics of the time series has not changed because the points are situated along the same line. So, there is no need to use a piecewise-linear regression model.

3.2. Time-series alignment

The very easy way to determine the fitted model to the time series data consists to use the command "Add Trendline" on the graph (Figure 26). The model equation and its R-squared statistic can be shown.

4. Verification

Figure 26 shows the original time series and two fitted models. For cubic polynomial $y = 1{,}3335t^3 + 7{,}281t^2 + 903{,}22t + 6\,613{,}7$ R-squared statistic is 0,9865. For linear function $y = 1\,576{,}7t + 3\,527{,}1$ R^2 statistic is 0,9612. If linear R^2 statistic is almost equal to nonlinear R^2 statistic, for forecasting linear regression model will be used.

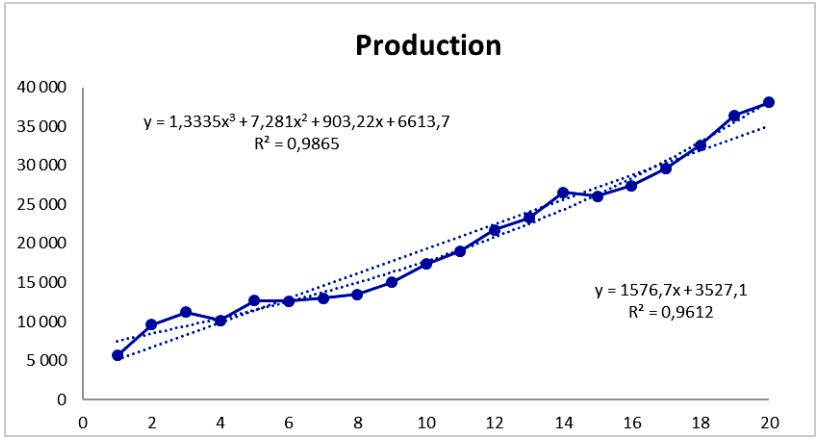


Figure 26 – **Time series with two fitted models**

5. Forecasting

In cell A23 enter 2011. In the cell B23 computer the forecasting production by linear function $y = 1576,7t + 3527,1$; to forecast t should be equal 22 (22 is the counting number of year 2011 if we count from 1997). Using linear regression model in 2011 the production of a product will be 38 215,2 rubles.

Questions for self-control

1. What is a time series?
2. Explain the meaning of trend effects, seasonal variations and random error.
3. What is the autocorrelation function?
4. How is the autocorrelation function used during the time series analysi?
5. What is the correlogram?
6. What is the general form of the multiplicative and additive time series model?
7. What is the purpose of analyzing the structure of seasonal time series of oscillations?
8. What tests are used to test hypotheses about the structural stability of the time series?
9. In any case violated the structural stability of the time series?

10. What is meant by the analytic alignment of the time series?

11. What are the most common known models used for the analytical smoothing of the time series?

12. What is meant by the linearized transformations? How are they used in the OLS?

13. How to evaluate the quality of the determined model?

14. How to obtain a point forecast for the time series model?

Individual task

The dynamics of output of some enterprise is characterized by the data presented in Table 14 (for individual task to the volume of output, we need to add the number $120 \cdot k$, where k is a student personal number in the students journal). Proceed as follows:

- analyze the structure of the time series;
- check the hypothesis of the structural stability of the series;
- carry out an analytical alignment of the time series;
- make a forecast for 2018;
- make a report.

Table 14 – **Original observations, rubles**

Year	Annual production	Year	Annual production
1997	5 665	2007	19 037
1998	9 570	2008	21 748
1999	11 172	2009	23 298
2000	10 150	2010	26 570
2001	12 704	2011	26 080
2002	12 588	2012	27 446
2003	13 018	2013	29 658
2004	13 471	2014	32 573
2005	15 017	2015	36 435
2006	17 356	2016	38 100

THEME 4. TIME SERIES WITH SEASONAL VARIATION

A time series that exhibits trend is a nonstationary time series. Modeling and forecasting of such a time series is greatly simplified if we can eliminate the trend. One way to do this is to fit a regression model describing the trend component to the data and then subtracting it out of the original observations, leaving a set of residuals that are free of trend.

Another approach to removing trend is by differencing the data; that is, applying the difference operator to the original time series to obtain a new time series, say,

$$x_t = y_t - y_{t-1} = \nabla y_t,$$

where ∇ is the (backward) difference operator.

Another way to write the differencing operator is in terms of a backshift operator *B*, defined as $By_t = y_{t-1}$, so

$$x_t = (1-B)y_t = \nabla y_t = y_t - y_{t-1},$$

with $\nabla = (1 - B)$.

Differencing can be performed successively if necessary until the trend is removes; for example, the second difference is

$$x_t = \nabla^2 y_t = \nabla(\nabla y_t) = (1-B)^2 y_t = (1-2B+B) = y_t - 2y_{t-1} + y_{t-2},$$

In general, powers of the backshift operator and the backward difference operator are defined as

$$B^d y_t = y_{t-d};$$

$$\nabla^d = (1-B)^d.$$

Differencing has two advantages relative to fitting a trend model to the data. First, it does require of any parameters, so it is a more parsimonious (i.e. simpler) approach; and second, model fitting assumes that the trend is fixed throughout the time series history and will remain so in the (at least

immediately) future. In other words, the trend component, once estimated, is assumed to be deterministic. Differencing can allow the trend component to change through time. The first difference accounts for a trend that impacts the change in the mean of the time series, the second difference accounts for changes in the slope of the time series, and so forth. Usually, one or two differences are all that is required in practice to remove an underlying trend in the data.

Seasonal, or both trend and seasonal, components are present in many time series. Differencing can also be used to eliminate seasonality. If s is the number of seasons (typically $s = 4$ or $s = 12$) the seasonal difference operator is define as

$$\nabla_d y_t = (1 - B^d) = y_t - y_{t-d}.$$

For example, if we had monthly data with an annual season (a very common situation), we would likely use $d = 12$, so the seasonally differenced data would be

$$\nabla_{12} y_t = (1 - B^{12}) = y_t - y_{t-12}.$$

When both trend and seasonal components are simultaneously present, we can sequentially difference to eliminate these effects. That is, first seasonally difference to remove seasonal component ant then difference one or times using the regular difference operator to remove the trend.

Formulation of the problem

The following is the given table which includes 16 quarters denoted as t and electricity consumption denoted as y_t (Table 15).

Table 15 – **Consumption of electricity by inhabitants of a given region, million kWh**

Quarters (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Electricity consumption (y_t)	6	4,4	5	9	7,2	4,8	6	10	8	5,6	6,4	11	9	6,6	7	10,8

Make a forecast of electricity consumption for the first half of the next year.

Computing technology in MS Excel to building time series seasonal variation and its econometrics analysis

Enter the data given in the Table 15. In the cell A1 enter the name of the first column "Quarter", t, and in the cell B1 enter the name of the second column "Electricity consumption". In the cells A2–A17 enter the data of the first row of the Table 15, and in the cells B2–B17 enter the data of the second row of the table.

Construct a graph of observations given in Table 15.

Examining the time series plot in Figure 27, there is both a strong positive trend as well as month-to-month variation. We can make a hypothesis that the time series contains three components: a random error, trend and seasonality. So the model should include both a trend and a seasonal component. It also appears that the magnitude of the seasonal variation does not vary with the level of the series, so an additive model is appropriate.

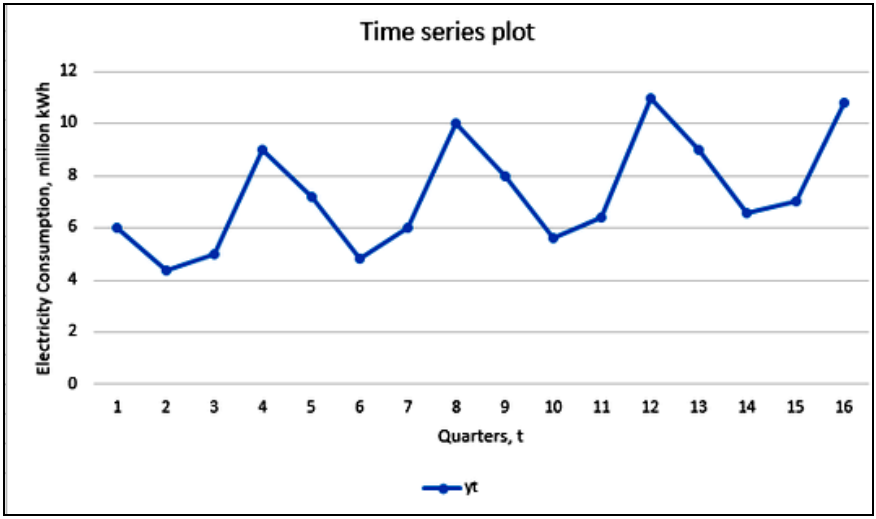


Figure 27 – **Time series plot of and electricity consumption**

We will for this case assume that the seasonal time series can be represented by the following model:

$$y_t = T_t + S_t + \varepsilon_t,$$

where T_t represents he level or linear trend component and can in turn be represented by $a + b \cdot t$;

S_t represents the seasonal adjustment with $S_t = S_{t+s} = S_{t+2s} = \ldots$ for $t = 1, \ldots, s-1$, where s is the length of the season (period) of the cycles;

ε_t are assumed to be uncorrelated with mean 0 and constant variance σ_ε^2.

Sometimes the level is called the permanent component. One usual restriction on this model is that the seasonal adjustments add to zero during one season

$$\sum_{t=1}^{s} S_t = 0.$$

When a trend and seasonal effect present in the time series, value of each subsequent level depends on previous values. Correlation between successive levels of the time series is determined by autocorrelation of levels.

To ensure statistical significance of autocorrelation coefficients, determine the number of periods for which the autocorrelation coefficient is calculated. The maximum lag should not be less than $n/4=16/4=4$. On the sheet "Given data" enter "Lag" in the cell D1. In the cells D2:D6 enter 1, 2, …, 9 using auto-complete feature.

In the cell E1 enter "Autocorrelation coefficient".

In the cell E2 enter the formula below to calculate autocorrelation coefficient r1

=CORREL(B2:B16,B3:B17).

In the cell E3 enter the formula below to calculate autocorrelation coefficient r2

=CORREL(B2:B15,B4:B17).

The rest of the autocorrelation coefficients are calculated as it is shown in the column E in Figure 28.

Cells in columns D:I fill as it is shown at Figure 28.

D	E	F	G	H	I
Lag	Autocorrelation coefficients	Sample size	Autocorrelation coefficients significance		t cr
1	=CORREL(B2:B16,B3:B17)	15	t1	=E2*SQRT((F2-2)/(1-E2^2))	=T.INV.2T(0.05, F2-2)
2	=CORREL(B2:B15,B4:B17)	14	t2	=E3*SQRT((F3-2)/(1-E3^2))	=T.INV.2T(0.05, F3-2)
3	=CORREL(B2:B14,B5:B17)	13	t3	=E4*SQRT((F4-2)/(1-E4^2))	=T.INV.2T(0.05, F4-2)
4	=CORREL(B2:B13,B6:B17)	12	t4	=E5*SQRT((F5-2)/(1-E5^2))	=T.INV.2T(0.05, F5-2)
5	=CORREL(B2:B12,B7:B17)	11	t5	=E6*SQRT((F6-2)/(1-E6^2))	=T.INV.2T(0.05, F6-2)
6	=CORREL(B2:B11,B8:B17)	10	t6	=E7*SQRT((F7-2)/(1-E7^2))	=T.INV.2T(0.05, F7-2)
7	=CORREL(B2:B10,B9:B17)	9	t7	=E8*SQRT((F8-2)/(1-E8^2))	=T.INV.2T(0.05, F8-2)
8	=CORREL(B2:B9,B10:B17)	8	t8	=E9*SQRT((F9-2)/(1-E9^2))	=T.INV.2T(0.05, F9-2)
9	=CORREL(B2:B8,B11:B17)	7	t9	=E10*SQRT((F10-2)/(1-E10^2))	=T.INV.2T(0.05, F10-2)
max	=MAX(E2:E10)				

Figure 28 – **Formulas to calculate autocorrelation coefficients and its critical values**

If the greatest autocorrelation coefficient is the autocorrelation coefficient with the lag $\tau > 1$, then the time series has seasonal effect with period τ, trend and random fluctuations. If $\tau = 1$ then time series contains only trend and random fluctuations.

The first autocorrelation coefficient ($\tau = 1$) is equal to 0,165 (Figure 29). It is not statistically significant because 0,6 < 2,1. Similarly the third autocorrelation coefficient ($\tau = 3$) is not significant. This indicates a weak depending of current levels y_t on the levels y_{t-1} и y_{t-3}. There is a marked dependence of a number of current levels y_t on levels y_{t-2}.

D	E	F	G	H	I
Lag	Autocorrelation coefficients	Sample size	Autocorrelation coefficients significance		t cr
1	0.165	15	t1	0.604	2.160
2	-0.567	14	t2	-2.384	2.179
3	0.114	13	t3	0.379	2.201
4	**0.983**	12	t4	16.943	2.228
5	0.119	11	t5	0.359	2.262
6	-0.722	10	t6	-2.952	2.306
7	-0.003	9	t7	-0.009	2.365
8	0.974	8	t8	10.499	2.447
9	0.097	7	t9	0.218	2.571
max	**0.983**				

Figure 29 – **Values of autocorrelation coefficients and their significance**

If the greatest autocorrelation coefficient is the fourth autocorrelation coefficient ($\tau = 4$) which is equal to 0,98, therefore the time series contains the seasonal component with a period of 4 quarters. As the value of the fourth autocorrelation coefficient is close to one, we can assume the presence of a linear trend in the time series (Figure 29).

Construct a correlogram. It is a graph of autocorrelation coefficients. So to construct a correlogram select cells E2:E10.

The wave-like pattern in Figure 30 suggests that the time series contains seasonality. Thus, the autocorrelation function values analysis suggests the presence of seasonally with the period of four quarters in the studied time, which is also seen in the correlogram.

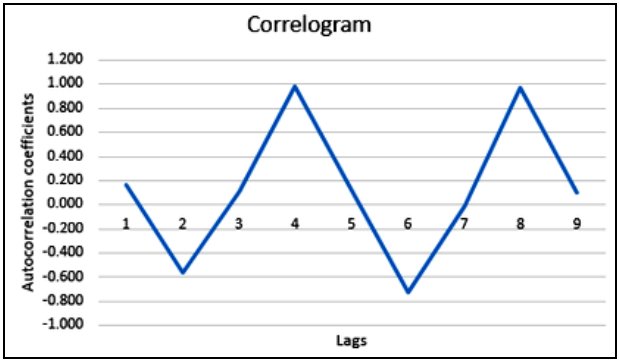


Figure 30 – **Graph of autocorrelation function**

As there are several components in the time series, it is necessary to determine how to combine them into models: to multiply or add. The time series plot is analyzed. The electricity consumption is greater in autumn and winter (quarters I and IV of each year) than in spring or summer (quarters II and III of each year). Analysis of Figure 27 shows that the amplitude of fluctuations of the time series is constant. Therefore, we can assume that the model is additive and represents the sum of the components. The seasonal time series is represented by the following formula

$$Y = T + S + E,$$

where T – represents the level or linear trend component and can it turn be represented by $a + b \cdot t$;

S – a seasonal component;

E – a random component.

To construct an additive model is to calculation the values of *T*, *S* and *E* for each level of the time series.

The procedure for constructing the model is as follows.

- *Step 1*. To smooth observations using a simple moving average.
- *Step 2*. To calculate a seasonal effect for each quarter in the data set.
- *Step 3*. To eliminate the seasonal component from time series data; in the additive model $Y - S = T + E$.
- *Step 4*. To fit an appropriate trend model to the data where only trend and random error are present ($T + E$).
- *Step 5*. To make calculation using additive model $T + S$.
- *Step 6*. To determine R-squared statistic.

Let's follow the steps above:

Step 1. It is useful to overlay a smoothed version of the original data on the time series plot to help reveal patterns in the original data. A simple moving average of span N is an average of N observations. In our case we will use N = 4. If we let $\tilde{y}_t$ be the moving average, then the 4-quater moving average at time period T is

$$\tilde{y}_T = \frac{y_T + y_{T-1} + y_{T-2} + y_{T-3}}{4} = \frac{1}{4}\sum_{t=T-3}^{T} y_t.$$

The moving average has less variability than the original observations (Figure 31).

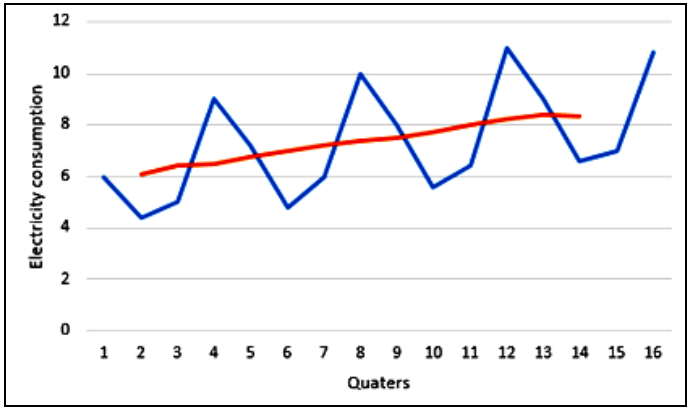


Figure 31 – **Time series plot with 4-quater moving average**

We will calculate smoothing data using weighted centered moving average

$$\tilde{y}_t^c = \frac{1}{8}y_{t-2} + \frac{1}{4}y_{t-1} + \frac{1}{4}y_t + \frac{1}{4}y_{t+1} + \frac{1}{8}y_{t+2}.$$

It is easier not to use the formula above, but to define $\tilde{y}_t^c$ as $\tilde{y}_t^c = \frac{1}{2}(\tilde{y}_t + \tilde{y}_{t-1})$ which is an average of the moving average $\tilde{y}_t$ at time period *t* and the moving average $\tilde{y}_{t-1}$ at time period *t* – 1, where

$$\tilde{y}_t = \frac{1}{4}(y_{t-1} + y_t + y_{t+1} + y_{t+2}).$$

Calculate smooth data as it is shown in column C and D at Figure 32.

	A	B	C	D	E
1	Quater, t	Electricity consumption, yt	Moving average	Centered moving average	Seasonality estimation
2	1	6			
3	2	4.4	=AVERAGE(B2:B5)		
4	3	5	=AVERAGE(B3:B6)	=AVERAGE(C3:C4)	=B4-D4
5	4	9	=AVERAGE(B4:B7)	=AVERAGE(C4:C5)	=B5-D5
6	5	7.2	=AVERAGE(B5:B8)	=AVERAGE(C5:C6)	=B6-D6
7	6	4.8	=AVERAGE(B6:B9)	=AVERAGE(C6:C7)	=B7-D7
8	7	6	=AVERAGE(B7:B10)	=AVERAGE(C7:C8)	=B8-D8
9	8	10	=AVERAGE(B8:B11)	=AVERAGE(C8:C9)	=B9-D9
10	9	8	=AVERAGE(B9:B12)	=AVERAGE(C9:C10)	=B10-D10
11	10	5.6	=AVERAGE(B10:B13)	=AVERAGE(C10:C11)	=B11-D11
12	11	6.4	=AVERAGE(B11:B14)	=AVERAGE(C11:C12)	=B12-D12
13	12	11	=AVERAGE(B12:B15)	=AVERAGE(C12:C13)	=B13-D13
14	13	9	=AVERAGE(B13:B16)	=AVERAGE(C13:C14)	=B14-D14
15	14	6.6	=AVERAGE(B14:B17)	=AVERAGE(C14:C15)	=B15-D15
16	15	7			
17	16	10.8			

Figure 32 – **Formulas to calculate the moving average, the centered moving average and seasonality estimation**

Step 2. Calculate seasonality estimations as it is shown at scheme in Figure 33.

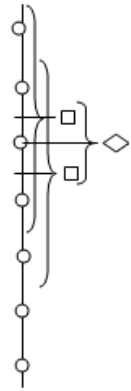


Figure 33 – **Scheme of seasonal adjustment**

Formulas to obtain seasonality estimations are presented at column E in Figure 32. The results are shown in Figure 34.

	A	B	C	D	E
1	Quater, t	Electricity consumption, yt	Moving average	Centered moving average	Seasonality estimation
2	1	6			
3	2	4.4	6.1		
4	3	5	6.4	6.25	-1.25
5	4	9	6.5	6.45	2.55
6	5	7.2	6.75	6.625	0.575
7	6	4.8	7	6.875	-2.075
8	7	6	7.2	7.1	-1.1
9	8	10	7.4	7.3	2.7
10	9	8	7.5	7.45	0.55
11	10	5.6	7.75	7.625	-2.025
12	11	6.4	8	7.875	-1.475
13	12	11	8.25	8.125	2.875
14	13	9	8.4	8.325	0.675
15	14	6.6	8.35	8.375	-1.775
16	15	7			
17	16	10.8			

Figure 34 – **Seasonality estimations**

Now a seasonal factor can be calculated for each period in the season using seasonality estimations at column E in Figure 33. The seasonal indices are computed by taking the average of all of the seasonal factors for each period in the season. In this example, all of the first quarter seasonal factors are average to produce a first quarter season index; all of the second quarter seasonal factors are average to produce a second quarter season index; and so on (Figure 35).

	A	E	F	G
1	Quater, t	Seasonality estimation	Quater	Average of the seasonal factors
2	1		1	=AVERAGE(E6,E10,E14)
3	2		2	=AVERAGE(E7,E11,E15)
4	3	=B4-D4	3	=AVERAGE(E4,E8,E12)
5	4	=B5-D5	4	=AVERAGE(E5,E9,E13)
6	5	=B6-D6	sum	=SUM(G2:G5)
7	6	=B7-D7		
8	7	=B8-D8		
9	8	=B9-D9		
10	9	=B10-D10		
11	10	=B11-D11		
12	11	=B12-D12		
13	12	=B13-D13		
14	13	=B14-D14		
15	14	=B15-D15		
16	15			
17	16			

Figure 35 – **Average of all of the seasonal factors for each quarter in the year**

The results are shown in Figure 36.

F	G
Quater	Average of the seasonal factors
1	0.6
2	-1.958
3	-1.275
4	2.708
sum	0.075

Figure 36 – **Average of the seasonal factors for each quarter**

It is usually assumed that in models with seasonal components, seasonal effects in the season are mutually compensating. In the additive model it is reflected in the fact that the sum of the all seasonal indices should be zero:

$$\sum_{t=1}^{4} S_t = 0.$$

For our example the sum of the all seasonal indexes is equal to 0,075 (the last line in Figure 36). The coefficient to adjust seasonal indexes is k = 0,01875 (Figures 37, 38).

F	G	H
Quater	Average of the seasonal factors	Adjusted seasonal index
1	=AVERAGE(E6,E10,E14)	=G2-G7
2	=AVERAGE(E7,E11,E15)	=G3-G7
3	=AVERAGE(E4,E8,E12)	=G4-G7
4	=AVERAGE(E5,E9,E13)	=G5-G7
sum	=SUM(G2:G5)	=SUM(H2:H5)
k	=AVERAGE(G2:G5)	

Figures 37 – **Determination of coefficient *k***

F	G	H
Quater	Average of the seasonal factors	Adjusted seasonal index
1	0.6	0.581
2	-1.958	-1.977
3	-1.275	-1.294
4	2.708	2.690
sum	0.075	0.000
k	0.01875	

Figures 38 – **Adjusted seasonal indexes for all quarters**

Adjusted seasonal index is the difference between the average of the seasonal factors for a quarter and the coefficient *k*. Adjusted seasonal in-

dexes are computed at column H, Figures 37, 38. Thus the sum of all adjusted seasonal indices is equal to zero.

Thus, the following seasonal indexes were obtained:

- for the first quarter: $S_1 = 0{,}581$;
- for the second quarter: $S_2 = -1{,}977$;
- for the third quarter: $S_3 = -1{,}294$;
- for the fourth quarter: $S_4 = 2{,}690$.

Step 3. Eliminate the seasonal component from original data by subtracting adjusted seasonal indexes from the original time series. We obtain the time series without the seasonal effect at the column M. It contains only trend and random component: $Y - S = T + E$. Figure 39 give formulas to operate.

J	K	L	M	N	O	P	Q
Quater, t	Electricity consumption, yt	Adjusted seasonal index	T+E=Y-S	Trend, T	T+S	E=Y-(T+S)	Y-Ya
1	6	0.581	5.419	5.902	6.483	-0.483	-1.3
2	4.4	-1.977	6.377	6.088	4.111	0.289	-2.9
3	5	-1.294	6.294	6.275	4.981	0.019	-2.3
4	9	2.690	6.310	6.461	9.151	-0.151	1.7
5	7.2	0.581	6.619	6.648	7.229	-0.029	-0.1
6	4.8	-1.977	6.777	6.834	4.857	-0.057	-2.5
7	6	-1.294	7.294	7.020	5.727	0.273	-1.3
8	10	2.690	7.310	7.207	9.896	0.104	2.7
9	8	0.581	7.419	7.393	7.974	0.026	0.7
10	5.6	-1.977	7.577	7.580	5.603	-0.003	-1.7
11	6.4	-1.294	7.694	7.766	6.472	-0.072	-0.9
12	11	2.690	8.310	7.952	10.642	0.358	3.7
13	9	0.581	8.419	8.139	8.720	0.280	1.7
14	6.6	-1.977	8.577	8.325	6.348	0.252	-0.7
15	7	-1.294	8.294	8.512	7.218	-0.218	-0.3
16	10.8	2.690	8.110	8.698	11.388	-0.588	3.5
					Verification	average, Ya	7.3
						numerator	1.098
						denominator	67.12
						R^2	0.9836

Figure 39 – **Values of additive model and its verification**

Step 4. Determine the trend component *T* of the additive model of time series as in theme 1. Put the regression statistics on another sheet with the title Regression. Regression statistics are represented on Figure 40.

1	SUMMARY OUTPUT								
2									
3	*Regression Statistics*								
4	Multiple R	0,956541002							
5	R Square	0,914970688							
6	Adjusted R Square	0,908897166							
7	Standard Error	0,280060809							
8	Observations	16							
9									
10	ANOVA								
11		*df*	*SS*	*MS*	*F*	*Significance F*			
12	Regression	1	11,81602042	11,81602	150,6491	6,997E-09			
13	Residual	14	1,098076797	0,078434					
14	Total	15	12,91409722						
15									
16		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
17	Intercept	5,715416667	0,146865127	38,91609	1,14E-15	5,4004223	6,03041104	5,400422296	6,030411037
18	Quarters	0,186421569	0,01518843	12,27392	7E-09	0,15384563	0,21899751	0,153845626	0,218997511

Figure 40 – **Statistics of linear regression**

Thus the trend component for electricity consumption is

$$T = 5{,}715 + 0{,}186 \cdot t.$$

The trend component for each observation is in the table “Residual output” (Figure 41).

22	RESIDUAL OUTPUT		
23			
24	*Observation*	*Predicted T+E=Y-S*	*Residuals*
25	1	5,90183824	-0,483088235
26	2	6,0882598	0,288823529
27	3	6,27468137	0,019068627
28	4	6,46110294	-0,150686275
29	5	6,64752451	-0,02877451
30	6	6,83394608	-0,056862745
31	7	7,02036765	0,273382353
32	8	7,20678922	0,103627451
33	9	7,39321078	0,025539216
34	10	7,57963235	-0,00254902
35	11	7,76605392	-0,072303922
36	12	7,95247549	0,357941176
37	13	8,13889706	0,279852941
38	14	8,32531863	0,251764706
39	15	8,5117402	-0,217990196
40	16	8,69816176	-0,587745098

Figure 41 – **Trend component and residuals for each observation**

Copy predicted values of trend on the sheet Steps in column N. The trend line is on Figure 42.

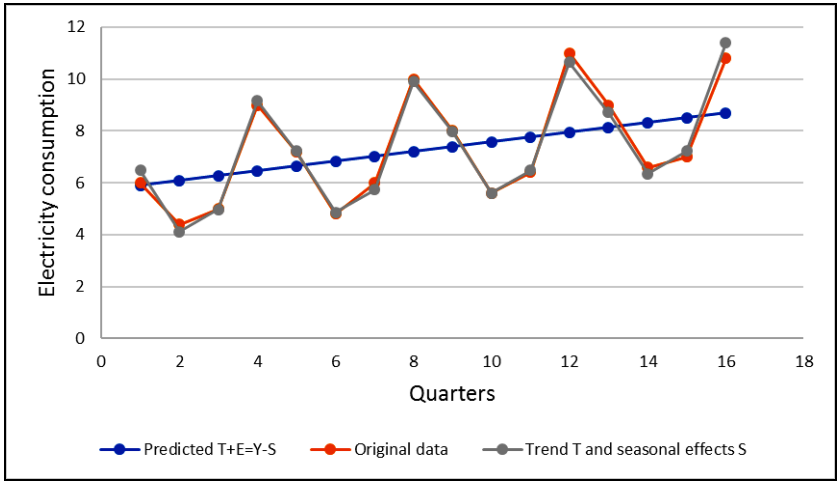


Figure 42 – **Electricity consumption in a region**

Step 5. Find out values for each level using additive model. For this, to every value in column N add corresponding seasonal index (as it shown in column O on Figure 39).

The graph of values ($T + S$) is represented on Figure 42. As we can see, the values calculated by additive model are near the time series observations.

Step 6 (model verification). The quality of the fitted model we will appreciate by R-squared statistic. Calculate $E = Y – (T + S)$ as the difference between the levels of the time series (column K) and the values obtained by the additive model (column O) (Figure 39). Enter the result in the column P. Calculate the difference between the levels of the time series and their average value (column Q, Figure 39). Determine the R-squared statistic for the fitted model as it is shown in Figure 43.

Verification	average, Ya	=AVERAGE(K2:K17)
	numerator	=SUMSQ(P2:P17)
	denominator	=SUMSQ(Q2:Q17)
	R^2	=1-Q20/Q21

Figure 43 – **Model verification**

Thus, the additive model explains 98,4% (see Figure 40) of the total variation of the levels of the time series of electricity consumption for the last 16 quarters.

Identification of the seasonal effect is made in the analysis of the structure of one-dimensional time series to forecast the values in future.

We want to forecast electricity consumption of the region during the first half of the next year.

Predictive value of the level of the time series in the additive model is the sum of the trend and seasonal components.

The volume of electricity consumed during the first half of the next year, i. e. the year fifth, is calculated as the sum of the volumes of electricity consumption in the I and II quarters of the fifth year (denoted respectively Y17 and Y18).

On the sheet "Regression" do the following (Figure 44):

- calculate trend components T17 and T18 by trend equation $T = 5{,}715 + + 0{,}186 \cdot t$;
- copy seasonal indexes for the first and second quarters from the sheet "Steps";
- add up the trend and seasonal components to forecast the electricity consumption in the first and second quarters;
- add up the electricity consumption in both quarters.

Forecasting					
T17	=B17+B18*17	S1	0.581	Y17	=F26+H26
T18	=B17+B18*18	S2	-1.977	Y18	=F27+H27
		First half of the next year			=SUM(J26:J27)

Figure 44 – **Formulas to forecast the electricity consumption**

Thus, the forecast of electricity consumption in the first half of the following (fifth) year will be 16,56 million kWh (Figure 45).

Forecasting					
T17	8.885	S1	0.581	Y17	9.466
T18	9.071	S2	-1.977	Y18	7.094
		First half of the next year			16.560

Figure 45 – **Forecast of electricity consumption**

Questions for self-control

1. What is a time series?
2. What are the main components of the time series?
3. What are the objectives of the analysis of time series?
4. Why is the autocorrelation function used?
5. How is the autocorrelation coefficient of the third-row?
6. What is the correlogram?
7. How to select the type of multiplicative and additive-term time-series models?
8. How to select the seasonal component of the time series?
9. Why is it needed to smooth the time series?
10. What are the methods of smoothing the time series?
11. What functions are used for analytical aligning of the time series?
12. How to make a forecast of the additive or multiplicative time series model?

Individual task

There are data on electricity consumption of the region for 16 quarters, presented in Table 16, where *k* is a student personal number in the students journal. Proceed as follows:

Proceed as follows:

- analyze the structure of the time series, check the hypothesis of the structural stability of the series;
- carry out an analytical alignment of the time series;
- make a forecast for the second half of the next year;
- make a report.

Table 16 – **Consumption of electricity by inhabitants of a given region, million kWh**

Quarters (*t*)	Electricity consumption (*yt*)
1	$6 + 0{,}2 \cdot k$
2	$4{,}4 + 0{,}2 \cdot k$
3	$5 + 0{,}2 \cdot k$
4	$9 + 0{,}2 \cdot k$
5	$7{,}2 + 0{,}2 \cdot k$
6	$4{,}8 + 0{,}2 \cdot k$

Table 16 (concluded)

Quarters (t)	Electricity consumption (yt)
7	$6 + 0{,}2 \cdot k$
8	$10 + 0{,}2 \cdot k$
9	$8 + 0{,}2 \cdot k$
10	$5{,}6 + 0{,}2 \cdot k$
11	$6{,}4 + 0{,}2 \cdot k$
12	$11 + 0{,}2 \cdot k$
13	$9 + 0{,}2 \cdot k$
14	$6{,}6 + 0{,}2 \cdot k$
15	$7 + 0{,}2 \cdot k$
16	$10{,}8 + 0{,}2 \cdot k$

BIBLIOGRAPHY

Adkins, L. Using gretl for Principles of Econometrics [Electronic resource]. – Mode of access : http://www.learneconometrics.com/gretl/using_gretl_for_POE4.pdf. – Date of access : 31.08.2017.

Books, C. Introductory Econometrics for Finance [Electronic resource]. – Mode of access : http://www.afriheritage.org/TTT/3%20Brooks_ Introductory%20Econometrics%20for%20Finance%20(2nd%20edition).pdf. – Date of access : 31.08.2017.

Creel, M. Graduate Econometrics Lecture Notes [Electronic resource]. – Mode of access : http://www.bseu.by/russian/faculty5/stat/docs/4/Creel, Graduate%20Econometrics.pdf. – Date of access : 15.08.2017.

Diebold, F. Econometrics. treamlined, Applied and e-Aware [Electronic resource]. – Mode of access : http://www.ssc.upenn.edu/~fdiebold/Teaching104/Econometrics.pdf. – Date of access : 12.08.2017.

Farnsworth, G. Econometrics in R [Electronic resource]. – Mode of access : http://cran.r-project.org/doc/contrib/Farnsworth-EconometricsInR.pdf. – Date of access : 25.07.2017.

Frisch, R. Editorial // Econometrica. – Vol. 1, No 1. – 1933. – P. 1–4.

Greene, W. Econometric Analysis [Electronic resource]. – Mode of access : http://stat.smmu.edu.cn/DOWNLOAD/ebook/econometric.pdf. – Date of access : 11.08.2017.

Gujarati, D. Basic Econometrics [Electronic resource]. – Mode of access : http://www.hse.ru/data/2011/04/26/1210823708/Gujarati%20D.N.%20Basic%20Econometrics,%203e,%201995.pdf. – Date of access : 12.07.2017.

Hansen, B. Econometrics [Electronic resource]. – Mode of access : http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf. – Date of access : 13.08.2017.

Hayashi, F. Econometrics [Electronic resource]. – Mode of access : http://press.princeton.edu/chapters/s6946.pdf. – Date of access : 31.08.2017.

Heij, C. Econometric Methods with Applications in Business and Economics [Electronic resource]. – Mode of access : http://www.listinet.com/bibliografia-comuna/Cdu339-A719.pdf. – Date of access : 17.07.2017.

LeSage, J. Applied Econometrics using MATLAB [Electronic resource]. – Mode of access : http://www.spatial-econometrics.com/html/mbook.pdf. – Date of access : 31.08.2017.

Montgomery, D. C. Introduction to time series analysis and forecasting / D. C. Montgomery, C. L. Jenningc, M. Kulahci. – New Jersey : Wiley, 2015. – 643 p.

Verbeek, M. A Guide to Modern Econometrics [Electronic resource]. – Mode of access : http://thenigerianprofessionalaccountant.files.wordpress.com/2013/04/modern-econometrics.pdf. – Date of access : 19.08.2017.

Wooldridge, J. Introductory Econometrics. A Modern Approach [Electronic resource]. – Mode of access : http://economics.ut.ac.ir/ documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf. – Date of access : 10.08.2017.

Авдашкова, Л. П. Эконометрика (продвинутый уровень) : пособие / Л. П. Авдашкова, М. А. Грибовская. – Гомель : Бел. торгово-экон. ун-т потребит. кооп., 2014. – 116 с.

CONTENT

Э 40 **Эконометрика** (продвинутый уровень) = Econometrics (advanced level) : пособие для реализации содержания образовательных программ высшего образования II ступени / авт.-сост. : Л. П. Авдашкова, М. А. Грибовская, С. В. Кравченко. – Гомель : учреждение образования "Белорусский торгово-экономический университет потребительской кооперации", 2018. – 72 с.

ISBN 978-985-540-451-5

В издании излагается технология построения и анализа эконометрических моделей. Расчеты выполняются с использованием MS Excel. Предназначено для использования при аудиторной и самостоятельной работе магистрантами экономических специальностей и всеми, кто интересуется эконометрикой.

УДК 519.862.6
ББК 65в63I

The publication sets out the technology of construction and analysis of econometric models. Calculations are performed using MS Excel. Designed for use in the classroom and independent work of a student of economics and all those interested in econometrics.

Учебное издание

Авдашкова Людмила Павловна
Грибовская Марал Атаевна
Кравченко Светлана Витальевна

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ)

Пособие
для реализации содержания образовательных программ высшего образования II ступени

На английском языке

Книга издана в авторской редакции

Технический редактор Т. В. Гавриленко
Компьютерная верстка Л. Ф. Барановская

Подписано в печать 18.04.18. Формат 60 × 84 $^{1}/_{16}$.
Бумага офсетная. Гарнитура Таймс. Ризография.
Усл. печ. л. 4,42. Уч.-изд. л. 3,73. Тираж 25 экз.
Заказ №

Издатель и полиграфическое исполнение:
учреждение образования "Белорусский торгово-экономический университет потребительской кооперации".
Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий
№ 1/138 от 08.01.2014.
Просп. Октября, 50, 246029, Гомель.
http://www.i-bteu.by

БЕЛКООПСОЮЗ
УЧРЕЖДЕНИЕ ОБРАЗОВАНИЯ
«БЕЛОРУССКИЙ ТОРГОВО-ЭКОНОМИЧЕСКИЙ
УНИВЕРСИТЕТ ПОТРЕБИТЕЛЬСКОЙ КООПЕРАЦИИ»

Кафедра информационно-вычислительных систем

ЭКОНОМЕТРИКА (ПРОДВИНУТЫЙ УРОВЕНЬ)

ECONOMETRICS (ADVANCED LEVEL)

Пособие
для реализации содержания образовательных программ
высшего образования II ступени

Гомель 2018