THE MINISTRY OF EDUCATION AND SCIENCE OF THE RUSSIAN FEDERATION
Federal State Autonomous Educational Institution of Higher Education
Lobachevsky State University of Nizhni Novgorod
National Research University

M.V. Kemaeva

ECONOMETRICS

TUTORIAL

Recommended by the Methodical Commission
of the Institute of Economics and Entrepreneurship, studying at the B.Sc.
Programme 38.03.01 “Economics” in English

Nizhni Novgorod

2017

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского»

М.В. Кемаева

ЭКОНОМЕТРИКА

Учебно-методическое пособие
по дисциплине «Эконометрика»

Рекомендовано методической комиссией Института экономики и предпринимательства ННГУ для иностранных студентов, обучающихся по направлению подготовки 38.03.01 «Экономика» (бакалавриат) на английском языке

Нижний Новгород
2017

УДК 330
ББК 65.05
К-35

К-35 **М.В. Кемаева**. Эконометрика: Учебно-методическое пособие. – Нижний Новгород: Нижегородский госуниверситет, 2017. − 38 с.

Рецензент: к.э.н., доцент Ю.А.Гриневич

В настоящем пособии изложены учебно-методические материалы по курсу «Количественные методы анализа экономики» для иностранных студентов, обучающихся в ННГУ по направлению подготовки 38.03.01 «Экономика» (бакалавриат).

Пособие дает возможность бакалаврам расширить основные знания о методах экономического анализа, которые объединяют экономическую теорию со статистическими и математическими методами анализа, овладевать умением комплексно подходить к вопросам экономического развития, использовать различные источники информации; развивать экономическое мышление

Ответственный за выпуск:
председатель методической комиссии ИЭП ННГУ,
к.э.н., доцент Летягина Е.Н.

УДК 330
ББК 65.05

INTRODUCTION

Modeling of economic processes is a really complicated procedures. Econometrics is part of the disciplines of mathematical modeling of the economy based on methods of statistical analysis. The main problem in econometrics is to build and estimate an econometric model from the point of view of its use for the description, analysis and prediction of real economic processes on the basis of factor approach.

Econometrics is intended to learn various methods of identifying and formalization relationships and regularities through econometric model based on statistical data. Within the framework of the course they study results, methods and techniques of economic theory, descriptive statistics and statistical mathematical tools to quantify really existing correlations. The students acquire necessary skills to formalize mastered dependencies and their use for description, analysis and forecasting of real economic processes.

The manual includes parts:

- theoretical material on the main themes of the course;
- tasks for independent work, cover the most important topics of this discipline and guidance on their completion of assignments;
- statistical tables necessary to perform control tasks.

In this tutorial presents only basic learning materials in the discipline of Econometrics. For a deeper understanding of the methods of econometric analysis, it is recommended to read additional literature.

CHAPTER 1. GENERAL INFORMATION

1.1 THE SUBJECT OF ECONOMETRICS

Econometrics is concerned with model building. An intriguing point to begin the inquiry is to consider the question, “What is the model?” The statement of a “model” typically begins with an observation or a proposition that one variable “is caused by” another, or “varies with another,” or some qualitative statement about a relationship between a variable and one or more covariates that are expected to be related to the interesting one in question. The model might make a broad statement about behavior, such as the suggestion that individuals’ usage of the health care system depends on, for example, perceived health status, demographics such as income, age, and education, and the amount and type of insurance they have. It might come in the form of a verbal proposition, or even a picture such as a flowchart or path diagram that suggests directions of influence. The econometric model rarely springs forth in full bloom as a set of equations. Rather, it begins with an idea of some kind of relationship. The natural next step for the econometrician is to translate that idea into a set of equations, with a notion that some feature of that set of equations will answer interesting questions about the variable of interest. To continue our example, a more definite statement of the relationship between insurance and health care demanded might be able to answer, how does health care system utilization depend on insurance coverage? Specifically, is the relationship “positive”—all else equal, is an insured consumer more likely to “demand more health care,” or is it “negative”? And, ultimately, one might be interested in a more precise statement, “how much more (or less)”? This and the next several chapters will build up the set of tools that model builders use to pursue questions such as theseusing data and econometric methods.

There is no single understanding of the term "econometrics" at the present time. We can only talk about the meaning of this term as "the science of economic measurement”. However, the essence of the subject "econometrics" is much broader. It can be defined as a method of economic analysis that integrates economic theory with statistical and mathematical methods of analysis. Economic theory usually provides answers to qualitative aspects. However, to make correct decisions, it is often necessary to quantify the characteristics of the phenomenon. For example, to answer the question: how will change in the volume of production, when the capital increase by a certain amount? [2]

This problem is solved in the framework of the econometric analysis, which is aimed to develop econometric models that use statistical methods to make specific quantitative relations of the overall quality of the laws, due to the economic theory. In other words, the main purpose of econometrics is modeling of specific quantitative description of the dependencies that really exist between various economic indicators for that purpose:

• to predict trends in economic and business processes for more effective and informed decisions;

• simulate different possible scenarios of socio - economic development of the analyzed system to determine how changes in particular manageable system parameters affect the study outcome indicators.

It should be noted that econometric models are different from other economic - mathematical models by the fact that their construction is based on the processing of real data and verifying their correctness on the base of statistical operations and criteria.

Thus, the econometrics can be considered as the science of building and analyzing the application of economic - mathematical models on the basis of statistical data to inform management and economic decisions.

1.2 THE RELATIONSHIP BETWEEN ECONOMIC INDICATORS

Regularities in the economy are formalized in the form of the relationship between economic indicators. It is possible to suppose that volume of production at a certain company Y depends on the cost of different types of resources $(x_1, x_2, ...x_k)$ and write: $Y = F(x_1, x_2, ..., x_k)$. Economic theory defines that the volume of demand for goods depends on the price and income level of the user K, i.e $Y = F(p, K)$. Each of these ratios is a model that determines how the variables are interrelated. In the General case, the relationship of the dependent variable Y and k independent $(x_1, x_2, ..., x_k)$ can be written: $Y = F(x_1, x_2, ..., x_k)$.

Independent variables are those that can be formalized by the researcher, whereas the dependent variables are variables that are measured using the generated dependencies. Independent variables in econometrics is also called the factor or explanatory variables.

If each set $\overline{X} = (x_1, x_2, ..., x_n)$ corresponds to one particular value of Y, then the communication is *functional*. A characteristic feature of functional relationships is that it is known the full list of factors determining the value of the dependent variable, as well as the exact mechanism of this effect, expressed by a specific equation in each case. Functional relationships actually exist in the economy (for example, the link between wage Y and output X with a simple piece-rate wages).

However, economic value added influenced by many factors, some of which act objectively (independently of the people will), others are the result of purposeful activity, in most cases. In addition, we often have to deal with incomplete information, studding economic dependencies. None know a complete list of factors affecting the analyzed indicator. Part of factors can be qualitatively heterogeneous and their influences realize ambiguous.

The value of the dependent variable is result of influence of a random variation in this case; they cannot be predicted surely, but with a certain probability. Such links are called *stochastic* and can be written as follows:

$$Y = F(x_1, x_2, ..., x_n) + \varepsilon \quad (1.1),$$

$x_i, \quad i = \overline{1, n}$ - independent (factor or explaining) variables;

$F(x_1, x_2,....x_n)$- part of dependent variable, formed under the influence of the considered factor(s) in a stochastic relation with Y;

ε - the part of the dependent variable formed by the action of uncontrolled, unaccounted factors, inaccurate measurement of factor variables and/or other random phenomena.

Thus, the main assumption is the requirement of *randomness of the values* when constructing econometric models.

The existence of dependence between the examined variables isn’t established in a mathematical way usually, but in the qualitative analysis of the phenomena that helps us to discover inner essence and the underlying causes and links.

The problem of econometric modeling is to establish the type of the function $F(x_1, x_2,..., x_n)$, the finding of such equation (estimation and estimator) that fits the nature studding interdependence in the best way.

1.3. TYPES OF ECONOMETRIC MODELS

There are many classifications of econometric models. For example, it is possible to differentiate regression model (factor variables and economic indicators) and time series (factor variables ordinal level time series).

Classification based on the number of equations identifies:

The regression model with one equation:

$Y = F(\bar{X}, a) + \varepsilon$,

$\bar{X} = (x_1, x_2,..., x_n)$ - factor variables, which can be any economic indicators;

$\bar{a} = (\bar{a}_1, a_2,..., a_k)$ - a vector of model parameters.

The regression model with one equation may have different functional forms. Most often built models of the following types:

- line $Y = a_0 + a_1x_1 + a_2x_2 + ... + a_kx_k$;
- power $Y = a_0 x_1^{a_1} x_2^{a_2} ... x_k^{a_k}$;
- polynomial $Y = a_0 + a_1x + a_2x^2 + ... + a_kx^k$ (usually, exponent is less than three).

It is useful to add that there are also hyperbolic, logarithmic, logistic, and other functions.

If the model contains only one factor variable (i.e. k=1) it is called a steam regression, when k>1 - multiple regression.

System of simultaneous equations.

These models are described by systems of equations. The system can consist of identities and regression equations, each of which may include dependent variables from other equations of the system (in addition to the independent factor variables). An example of such models is the model of supply and demand:

$$\begin{cases} Q_t^s = \alpha_1 + \alpha_2 \cdot P_t + \alpha_3 \cdot P_{t-1} + \varepsilon_t \\ Q_t^d = \beta_1 + \beta_2 \cdot P_t + \beta_3 \cdot Y_t + u_t \\ Q_t^s = Q_t^d \ , \end{cases}$$

Q_t^d - the demand for goods at time (moment) t;

Q_t^s - the offer of goods at time t;

P_t - the price of a commodity at time t;

Y_t - income at time t.

Model of income generation:

$$\begin{cases} C_t = \beta_0 + \beta_1 \cdot Y_t + \varepsilon_t \\ Y_t = C_t + I_t \end{cases}$$

Y_t - gross output;

C_t – consumption;

I_t – investment.

Classification based on the number of factors identifies: simple regression and multiple (more than 1 factor variable).

Multiple regression analysis is an extension of simple regression analysis to cover cases in which the dependent variable is hypothesized to depend on more than one explanatory variable. Much of the analysis will be a straightforward extension of the simple regression model, but we will encounter two new problems. First, when evaluating the influence of a given explanatory variable on the dependent variable, we now have to face the problem of discriminating between its effects and the effects of the other explanatory variables. Second, we shall have to tackle the problem of model specification. Frequently a number of variables might be thought to influence the behavior of the dependent variable; however, they might be irrelevant. We shall have to decide which should be included in the regression equation and which should be excluded. [1]

1.4. A METHOD OF CONSTRUCTING ECONOMETRIC MODELS

The process of designing and analyzing of econometric models is quite complicated and can be divided into the following main stages: *specification, identification and verification*.

The specification is based on economic theory, specialized knowledge and intuition of researcher about the analyzed economic system. In turn, the model specification includes: problem formulation and forms of communication choice.

Statement of the problem is the definition and formulation of the goals of modeling (selection dependent variable) and a set of factor variables. Usually, models include only the main, most significant factors that have a meaningful impact on the learning process. This principle is one of the main principles of modeling.

The second step consists of selection functional form of the link between the selected variables (specifications) $Y = f(\overline{X}, \overline{a})$ ($\overline{a} = (a_1, a_2, ..., a_k)$ is the vector of model parameters, which do not have specific numerical values at the stage of research).

Qualitative analysis of the phenomenon, knowledge of economic theory may suggest a particular functional form of communication. An important role is played by the analysis of the available statistical information: graphics, calculation of growth rates (when the construction of trend models) and so on and so forth.

It is possible to use some specific techniques selecting equation, when considering a particular class of statistical models (production functions, demand functions, etc.). For example, building a production function based on the source data it is usually possible to draw conclusions in relation to General changes in the value of product Y and such indicators as average and marginal product, the rate of substitution and others. Relevant characteristics is known for various forms of the equations. Thus, the prerequisites are formed for choosing informed and meaningful equation (s) of econometric models.

The *identification* of model is a statistical estimation of unknown model parameters. After the form of the equation have been selected, you can proceed to the calculation of the model parameters $\bar{a} = (a_1, a_2, ..., a_k)$.

Lets consider the regression model with one equation. The estimated equation model is $f(\bar{X}, \bar{a})$. The objective of the stage is to select function from the parametric family of functions $Y = f(\bar{X})$, $f(\bar{X}, \bar{a})$ describes the dependence of the observed Y values from the observed values $\bar{X}(x_1, x_2, ..., x_n)$ in the “best” way . To find the function is to choose the "best" estimators (coordinates of vector $\bar{a}$) in this case.

There are different methods of calculating the parameters of econometric models: the classical method of least squares (LSM), generalized LSV, etc.

Verification of model, as well as problem identification, is specifically associated with the construction of the econometric model. The actual construction of the model completes its identification. When identification is completed, some questions should be arisen:

- Solving of the problem of specification and identification of the model have been made successfully, i.e. whether it is possible to conclude that the use of models for forecasting and other calculations will bring quite adequate results, isn’t it?
- What is the accuracy of the forecast and simulation calculations based on the constructed model?

The essence of the problem of verification of econometric models is getting answers to these questions. Verification methods based on statistical procedures for testing hypotheses and statistical analysis of the accuracy characteristics for different methods for statistical estimation.

It should also be noted that the retrospective calculation principle used in the verification of econometric models. The essence of the principle is as follows: source of statistical data is divided into two parts: a training sample, including some observations, and examining sample, comprising the remaining portion of the original data. Next are the stages of specification and identification of the training sample. In the resulting model substituted exogenous variables from examining a sample and get the model values (retrospectively forecast) endogenous variables. Comparison of these model values with the corresponding actual values of examining sample allows us to analyze the adequacy of the model conclusions of reality and their accuracy.

1.5. THE SOURCE DATA FOR BUILDING ECONOMETRIC MODELS

Mathematical statistics base on the concept of general totality and sampling (some time, random). A *general totality* is the totality of all conceivable observations (or all conceivable objects applying to a specific type), which could be taken with the real complex of conditions [1]. The concept of the General totality is a mathematical abstraction. In practical work researchers deal with samples from the General totality. Sample is a limited set of objects of the General population, which can be viewed as an empirical analogue of the General totality.

The main requirement to the sample, its representativeness, i.e. the completeness and adequacy of representation of general totality to the interest of a researcher properties of the general totality. If the sample is defined incorrectly, researcher will take incorrect results (constructed model will not correspond to the real process and give wrong conclusions).

For example, let's study the regional economy. Demand for some goods depends on income. Researcher have included in the sample only families with high income. Obviously, he gets incorrect results because of non-representative sample.

If you consider the amount of income as a random variable, then the sample will be representative in the case if the corresponding relative frequency of sampling is about general totality corresponding relative frequency. Essential step is preparation and selection of statistical data. They should be agreed between themselves and have a uniform methodological basis of collection.

Even if you deal with the representative sample, it must present sufficiently large set of statistical observations, where each observation is characterized by numerical values of each factor $\overline{X} = (x_1, x_2, ..., x_n)$and the dependent variable Y. It is believed that the number of observations should be more then in 5-6 times greater than the number of parameters of the equation. Increasing sample size generally leads to higher reliability of the results of econometric studies.

Statistical data can be classified in two types: experimental and observational. The data of the first type get as a result of experiment.

Observational data are based on statistical reports and surveys. In econometric studies are primarily used observational statistical data, which are usually divided into two types: cross-sectional data (spatial) and time series.

Cross-sectional data are collected measuring any of the economic indicators for different entities (firms, households, etc. in one time.

Time series data for the same object at different points in time. The same dependence can be studied on the basis of the cross, and temporary data. For example, the production function of the industry, expressing the dependence of the volume of production from labor costs and production assets, can be obtained in two ways: based on data for one year in various enterprises (cross monitoring), or a few years of experience in the industry (time series).

Often the original statistical aggregate is formed from the combined cross-temporal data (panel data), for example, data from a number of companies over several reporting periods. There are a variety of methods to conduct data collection: questionnaires, direct observation, using of companies and firms internal reports, data, publications, statistical reports, etc.

1.6 TIME SERIES MODELING

By time series we mean a series of numerical values obtained at regular intervals. Specific values of this series are called the *levels of the series*. For example, time series are:

- series of daily observations of the price for goods over some period of time;
- daily volumes of the production of goods;
- monthly inflation or the consumer price index;
- quarterly estimates of gross national product or average wages (accepted in Russia for the quarterly indexing of pensions);
- annual data on the volume of company`s revenues and profits.

Time series, of course, is not limited to economic quantities. It is known for its use in the analysis of processes in power systems, nuclear power industry, chemical and petrochemical industries. In this case, it often uses smaller discreteness of time than in the economy, minutes and even seconds of processing the data on high-speed processes in the nuclear industry or in the study of transient processes in chemical kinetics.

The fundamental Statute, which is based on the use of time series prediction, is that the factors affecting the values of the numeric indicators of the system under study acted in some way in the past and do so at present. So, it is expected that they will act in a similar way in the near future. Therefore, the main purpose of time series analysis will be the assessment and identification of these uncontrollable factors aimed at predicting the future behavior of the system and the development of rational management decisions.

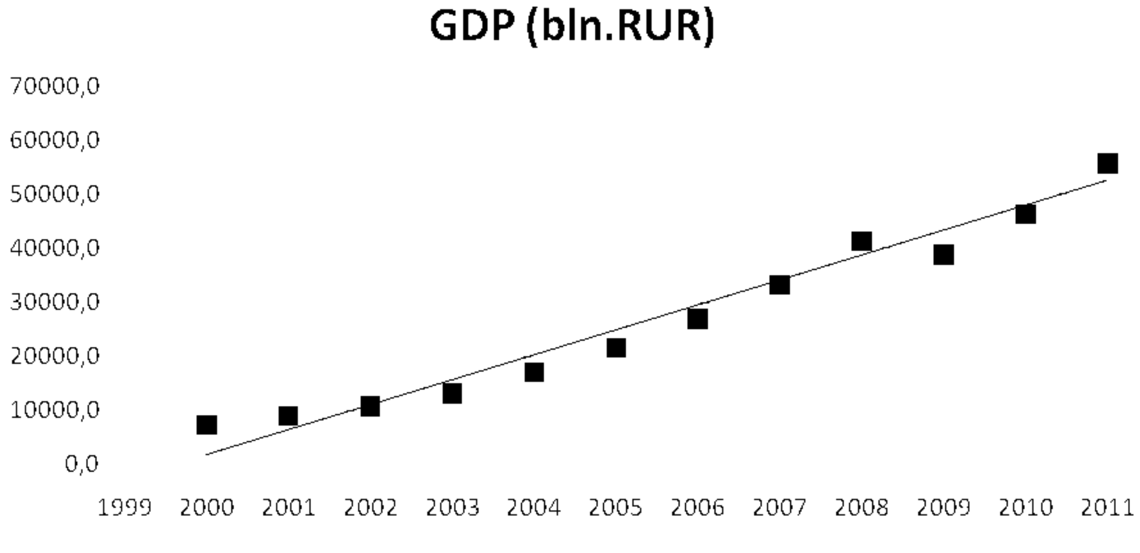

Figure 1.1 GDP Of Russian Federation 1999-2011.

Generally, we can say that the volume of GDP shows a marked tendency to growth for the specified period, and this common tendency (upward or downward) is called a trend. The points are used to show real data, the line - for the tendency.

The trend, however, is not the only component of the time series. Its levels can be distinguished periodically from fast to slow growth against the background of a clear increase. It is believed that the trend is complicated by the existence of cyclic components and random components. Analyzing the series with a shorter step (quarterly or monthly data) is targeted to short-time deviations from the trend, repeating regularly with some resistance. The deviations are explained by the existence of the seasonal component in a number of levels.

Cyclical component explains the deviations from the trend at intervals of 2 to 10 years. Usually this component may vary according to the length of the period and its intensity, and it may correlate with the business cycle. At the stage of rising business activity the values of the time series are above the trend, while in periods of recession and stagnation they are significantly lower than expected according to the trend.

Seasonal component determines short-term fluctuations due to changes in the activity within a year and can be found at a more or less fixed intervals time. It can be naturally tracked quarterly, monthly or more frequently.

It is natural to associate the seasonal component with the influence of traditions (seasonal and Christmas sales), social habits (higher activity in the resort business in summer and its decline in other seasons), religious factors (Christmas and Easter and other holiday and evens), weather (sale of ice-cream and soft drinks, activities at ski resorts).

The stochastic component causes deviations from the levels determined by trend, cycle and seasonal ingredients. It may be considered as random, and therefore, unpredictable; in terms of statistics it can be considered as an error of monitoring. It should be processed in the same way as random errors of measurements in statistics. It usually results from random phenomena of the external world - hurricanes, floods, strikes, the impact of political processes, such as elections, or the uncertainty of the outcome, turnover and uprising.

The only requirement to the data of a time series is that the measurement of the levels of the numeric values be spaced at uniform time intervals. The majority of economic indicators are published once a year. If the data is/are received at smaller intervals, quarterly or monthly, the procedure remains the same, except time series changing regularly over the whole period of observations (because of the seasonal component).

If a time series includes trend and seasonal changes, the value of each subsequent level will depend on the values of the previous.

The correlation between successive levels of time series are called *autocorrelation* levels of time series. It can be estimated by the index of correlation between the levels of the original time series and the levels (shifted) a few steps back (in time). This shift back in time, which is used to calculate correlation index, is called lag.

So if the correlation coefficient is calculated between the levels of yt and yt-1 (t=2,...,n), the lag equals 1. We denote the ratio of r1. It can be calculated by the formula:

$$r_1 = \frac{\sum\limits_{i=2}^{n}(y_t - \bar{y}_t)(y_{t-1} - \bar{y}_{t-1})}{\sqrt{\sum\limits_{i=2}^{n}(y_t - \bar{y}_t)^2 * \sum\limits_{i=1}^{n}(y_{t-1} - \bar{y}_{t-1})^2}},$$

$$\bar{y}_t = \frac{\sum\limits_{i=2}^{n} y_t}{n-1}, \qquad \bar{y}_{t-1} = \frac{\sum\limits_{i=2}^{n} y_{t-1}}{n-1}.$$

If the correlation coefficient is calculated between the levels of yt and yt-k t=k+1,...,n), the lag equals k, we denote the coefficient of rk. It can be calculated by the formula:

$$r_k = \frac{\sum\limits_{i=k}^{n}(y_t - \bar{y}_t)(y_{t-k} - \bar{y}_{t-k})}{\sqrt{\sum\limits_{i=k+1}^{n}(y_t - \bar{y}_t)^2 * \sum\limits_{i=k+1}^{n}(y_{t-k} - \bar{y}_{t-k})^2}}$$

(7.1.)

$$\bar{y}_t = \frac{\sum\limits_{i=\text{л}+1}^{n} y_t}{n-1}, \qquad \bar{y}_{t-1} = \frac{\sum\limits_{i=k+1}^{n} y_{t-1}}{n-1}.$$

CHAPTER 2. SIMPLE LINEAR REGRESSION

2.1. MODEL OF LINEAR REGRESSION

If the regression econometric model has one explanatory (factor) variable , it is called a model simple regression and it is possible to write:

$$Y = F(X,\overline{\alpha}) + \varepsilon, \quad (2.1)$$

$\overline{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_k)$ - the parameters of the model,

ε - the stochastic perturbation.

The function $F(X,\overline{\alpha})$ describing the link in variables can be linear or nonlinear. It is useful to analyze graphics to define a type of correlation (linear or nonlinear) in the case of simple regression. Graphical depiction of real statistical data as points in the Cartesian coordinate system is called the flowchart or path diagram (Fig. 1.1)

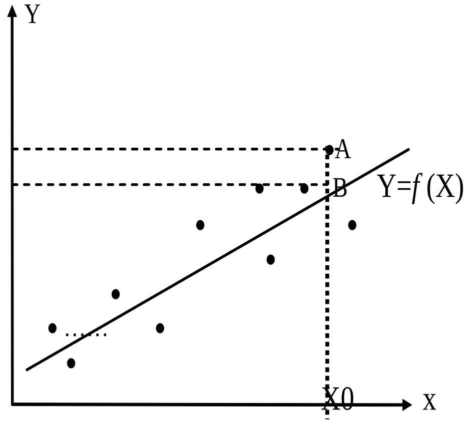

Fig. 1.1 Flowchart or path diagram

The correlation coefficient may indicate that two variables are associated with one another, but it does not give any idea of the kind of relationship involved. We will now take the investigation a step further in those cases for which we are willing to hypothesize than one variable depends on another. It must be stated immediately that one would not expect to find an exact relationship between any two economic variables, unless it is true as a matter of definition. In textbook expositions of economic theory, the usual way of dealing with this awkward fact is to write down the relationship as if it were exact and to warn the reader that it is really only an approximation. In statistical analysis, however, one generally acknowledges the fact that the relationship is not exact by explicitly including in it a random factor known as the disturbance term. We shall start with the simplest possible model [1]:

$Y = \alpha \cdot + \beta \cdot X + \varepsilon$

Based on visual analysis of Fig. 1.1 we can assume that the relationship between variables is close to linear and first function to analyze should become linear: $Y = a + b \cdot X$.

The linear regression model is the most common (and easy) view of the relationship between economic indicators. It usually serves as the starting point for the econometric analysis.

If we look at fig. 1.1, we will see disturbance between real point and estimated line. Why does the disturbance term exist? There are several reasons.

1. Omission of explanatory variables: The relationship between Y and X is almost certain to be a simplification. In reality there will be other factors affecting Y that have been left out of (fig. 1.1),and their influence will cause the points to lie off the line. It often happens that there are variables that you would like to include in the regression equation but cannot because you arebunable to measure them. For example, later on in this chapter we will fit an earnings function relating hourly earnings to years of schooling. We know very well that schooling is not the only determinant of earnings and eventually we will improve the model by including other variables, such as years of work experience. However, even the best-specified earnings function accounts for at most half of the variation in earnings. Many other factors affect the chances of obtaining a good job, like the unmeasurable attributes of an individual, and even pure luck in the sense of the individual finding a job which is a good match for his or her attributes. All of these other factors contribute to the disturbance term.

2. Aggregation of variables: In many cases the relationship is an attempt to summarize in aggregate a number of microeconomic relationships. For example, the aggregate consumption function is an attempt to summarize a set of individual expenditure decisions. Since the individual relationships are likely to have different parameters, any attempt to relate aggregate expenditure to aggregate income can only be an approximation. The discrepancy is attributed
to the disturbance term.

3. Model misspecification: The model may be misspecified in terms of its structure. Just to give one of the many possible examples, if the relationship refers to time series data, the value of Y may depend not on the actual value of X but on the value that had been anticipated in the previous period. If the anticipated and actual values are closely related, there will appear to be a relationship between Y and X, but it will only be an approximation, and again the disturbance term will pick up the discrepancy.

4. Functional misspecification: The functional relationship between Y and X may be misspecified mathematically. For example, the true relationship may be nonlinear instead of linear. Obviously, one should try to avoid this problem by using an appropriate mathematical specification, but even the most sophisticated specification is likely to be only an approximation, and the discrepancy contributes to the disturbance term.

5. Measurement error: If the measurement of one or more of the variables in the relationship is subject to error, the observed values will not appear to conform to an exact relationship, and the discrepancy contributes to the disturbance term.

The disturbance term is the collective outcome of all these factors. Obviously, if you were concerned only with measuring the effect of X on Y, it would be much more convenient if the disturbance term did not exist. However, in fact, part of each change in Y is due to a change in ε, and this makes life more difficult. For this reason, ε is sometimes described as noise.[1]

2.2. ESTIMATION OF SIMPLE LINEAR MODEL PARAMETERS. THE LEAST SQUARES METHOD

Lets suppose that there is a linear relationship between the explained variable and the explanatory variable (factor). This relationship can be written:

$$Y = \alpha \cdot + \beta \cdot X + \varepsilon, \text{ (2.2)}$$

α and β are the true model parameters, which could be obtained analyzing the General totality.

Then $\varepsilon_i = y_i - (\alpha + \beta \cdot X_i)$ is an error in the observation number i. However, even a when linear relation actual exists, the parameters α and β of this relationship remain unknown, and we can appreciate approximately their true values and estimate their values on base of the limited number of available sampling data.

Different lines can be drawn through the point at the flowchart, which parameters will be quite differing. We want to have such line $Y = a + b \cdot X$, which is the best among all straight lines in a certain sense, i.e. the "nearest" to the observation points in their entirety. It is necessary to define the concept of the closest linear function to any set of points in the plane. Measures of such closest location may be different. However, it should be related with the distance between the observation points and the line, i.e., the value

$$e_i = Y_i - (a + b \cdot X_i) = Y_i - \widehat{Y}_i, \text{ (2.3)}$$

i – numder or observation, $i = \overline{1, n}$.

The value e_i is called the error in the i-th observation (cuts AB in Fig. 2.1), $\hat{Y}$ - a theoretical value obtained by substitution in the equation models the observed values of factor variables (model estimation).

As a rule, errors are differing from zero, and some of them have a positive sign, other – negative. It depends on location of real point. If a particular point is above model line, e_i takes a positive value. If a particular point is below model line, e_i takes a negative value. So if researcher takes a total distance (sum of residuals $\sum_{i=1}^{n} e_i$) as measure to consider, it may be equal to zero.

If all deviations are squared and summed, the sum is non-negative and its magnitude will depend on the scatter of points around the line function. Different values of parameters determining different lines, and they will meet various sums of squared deviations:

$$U(\bar{a}) = \sum_{i=1}^{n} (Y_i - (a \ + bX_i))^2 = \sum_{i=1}^{n} e_i^{\ 2}. \text{ (2.4)}$$

The principle of least squares is the choice of these parameters a and b, which define the minimal meaning of function $U(\bar{a})$. The obtained estimates a and b of parameter α and β are called *the least squares estimates*.

Thus, LSM is a method of estimating the parameters of linear econometric models based on minimizing the sum of squared deviations of the observed values of the dependent variable Y_i from the desired linear function $\widehat{Y_i} = F(X_i, \bar{a})$.

Since the function is continuous, convex and bounded from zero below, it has the minimum value.The calculation of the parameters by LSM is a well-known mathematical problem of finding the minimum of the function. This minimal point can be found by equating to zero the partial derivatives of the function $U(\bar{a})$ with esteems. Let's write down the necessary conditions for an extremum:

$$\begin{cases} \dfrac{\partial U}{\partial a} = -2\sum_{i=1}^{n}(Y_i - a - b\cdot x_i) = 0 \\ \dfrac{\partial U}{\partial b} = -2\sum_{i=1}^{n} x_i \cdot (Y_i - a - b\cdot x_i) = 0 \end{cases} \quad (2.5)$$

We obtain a system of equations for calculation of LSM parameters. This system is called the system of normal equations.

Discover brackets and get the standard form of the normal equations:

$$\begin{cases} a\cdot n + b\cdot \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ a\cdot \sum_{i=1}^{n} X_i + b\cdot \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \end{cases} \quad (2.6)$$

Divide each the equations into n and obtain:

$$\begin{cases} a + b\cdot \overline{X} = \overline{Y} \\ a\cdot \overline{X} + b\cdot \overline{X^2} = \overline{X\cdot Y} \end{cases} \quad (2.7),$$

$\overline{X}$, $\overline{Y}$ etc. - average level of variable $(\overline{X} = \dfrac{\sum_{i=1}^{n} X_i}{n})$.

The first equation in (2.7) shows that the model line goes through the point of averages of observable sampling $(\overline{X}, \overline{Y})$.

We assume that among the observed values of X are not all numbers are equal, then $\overline{X^2} - \overline{X}^2 = \Delta \neq 0$ (the determinant of the system), therefore, the solution can be found by the Kramer rule:

$$\begin{cases} b = \dfrac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2} \\ a = \overline{Y} - b\cdot\overline{X} \end{cases} \quad (2.8)$$

The numerator in the formula for calculation of the parameter is the ratio of the covariance of variables X and Y:

$$Cov(X,Y) = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})\cdot(Y_k - \overline{Y}) = \overline{XY} - \overline{X}\cdot\overline{XУ},$$

and the denominator of the sample variance of X : $D(X) = \overline{X^2} - \overline{X}^2$.

Therefore, we can write (2.9)

$$b = \frac{Cov(X,Y)}{D(X)} \qquad (2.9)$$

One thing that should be accepted from the beginning is that you can never discover the true values of and, however much care you take in drawing the line. Parameters a and b are only estimates, and they may be good or bad. Once in a while your estimates may be absolutely accurate, but this can only be by coincidence, and even then you will have no way of knowing that you have hit the target exactly. This remains the case even when you use more sophisticated techniques. Drawing a regression line by eye is all very well, but it leaves a lot to subjective judgment. Furthermore, as will become obvious, it is not even possible when you have a variable *Y* depending on two or more explanatory variables instead of only one.[1]

2.3. THE CLASSICAL NORMAL LENEAR MODEL OF REGRESSION

The least-squares method allows to determine the estimates of the model parameters a and b. But it isn't known how close the values of the parameters to their theoretical counterparts α and β. Is it reliable the estimates obtained? To answer this question, more research is needed.

The properties of Parameters a and b can be judged only if we imposed certain conditions on the random member ε

There are the following four assumptions in the classical normal linear regression model (*Gauss-Markov conditions*):

1. The mathematical expectation of the random deviations is equal to zero for all observations: $M(\varepsilon_i) = 0$, i = 1,2,...,n,

This condition means that the random variation in average has no effect on the dependent variable. Each observation is characterized by a random member, that can be either positive or negative, but it should have no systematic bias.

2. $M(Y|X = X_i) = \alpha + \beta \cdot X_i$ - the second condition, which is automatically executed when the condition 1.

3. The dispersion of a random deviation constant for all observations:

$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ for i = 1,2,...,n,

and its value is unknown (one of the objectives of regression analysis is to estimate the dispersion).

This condition implies that, despite the fact that each observation is a random deviation may be different, the probability that the value ε will take some of this (positive or negative) value will be the same for all observations. The feasibility of this assumption is called homoscedasticity (constancy of the dispersion of the variance), and the impracticability - heteroscedasticity (variability of the dispersion of the variance).

4. Random deviations in all observations should be independent from each other, i.e. there is no systematic relationship between any random deviations. This condition, in particular, means that:

$$Cov(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$$

Random deviations must be distributed independently of the explanatory variables. So, if X - random value, then this condition is automatically performed and

$M(X_i \cdot \varepsilon_i) = 0$.

If conditions 1-4 are met, then estimates are unbiased, consistent and effective, (made by the method of LSM).

To verify the above properties, there are special statistical criteria.

In addition to the Gauss–Markov conditions, one usually assumes that the disturbance term is normally distributed. You should know all about the normal distribution from your introductory statistics course. The reason is that if u is normally distributed, so will be the regression coefficients, and this will be useful to us later in the chapter when we come to the business of performing tests of hypotheses and constructing confidence intervals for α and β using the regression results. The justification for the assumption depends on the Central Limit Theorem. In essence, this states that, if a random variable is the composite result of the effects of a large number of other random variables, it will have an approximately normal distribution even if its components do not, provided that none of them is dominant. The disturbance term u is composed of a number of factors not appearing explicitly in the regression equation so, even if we know nothing about the distribution of these factors (or even their identity), we are entitled to assume that they are normally distributed. [1]

These properties don't depend on a particular distribution type of values ε_i. However, it is generally assumed that they are distributed normally. This premise is necessary to check the statistical significance of the found estimates and determination of confidence intervals.

It is known that a linear combination of normally distributed random variables has a normal distribution. Therefore, the parameters a and b also have a normal distribution.

2.4 THE CONCEPT OF STATISTICAL SIGNIFICANCE

As already noted, the econometric model is usually based on the selected statistical data. Equation parameters, correlation coefficients and other features of the model are defined on the basis of sample observations. They will obviously be different from the corresponding values calculated for the General totality.

Therefore, the sample characteristics can be attributed to some errors related to incomplete coverage of observations of all units of the General totality. In turn, this requires verification of the reliability and statistical significance of parameters and other characteristics of the model. If you don't conduct such checks, it is possible to lead to false conclusions about the existence of a connection where none exists.

The *statistical significance of the result* is an estimated measure of confidence in its "truth" (in the sense of "representative sample").

To characterize the statistical significance should be used the concept of the level of statistical significance α. This indicator is in the decreasing dependence of the reliability of the result. Higher level α corresponds to a lower level of confidence to parameters and other characteristics of the model found in the base of a sample.

That α level represents the probability of error associated with the distribution of the observed results on the General totality. For example, α= 0,05 shows that there is a 5% probability that found the relationship between variables is only an incidental feature of the sample used. The choice of a certain significance level (above which the results are rejected as false) is arbitrary. Usually level α=0,05 is an acceptable level of statistical significance. However, it should be remembered that this level still includes quite large probability of error (5%). The results, significant level α= 0,1, are usually considered as statistically significant, and the results with the level $\alpha \le 0{,}05$or $\alpha \le$ 0,01 as highly significant.

Tests of statistical significance are carried out on statistical testing of hypotheses using t-statistics. Hypotheses about the statistical significance of some magnitude U are formulated as follows:

$H_0 : U = 0$

$H_1 : U \neq 0$

To verify the hypothesis is made:

$$t_U = \frac{|U - 0|}{s_U} = \frac{|U|}{s_U}, \text{ (2.10),}$$

t_u is called t – statistics, s_u is the standard error in estimation of magnitude U.

This attitude has a t-student distribution with n-2 degrees of freedom. The t-distribution is summarized into a theoretical tables depending on the chosen level of statistical significance and degree of freedom. The level of statistical significance chose by the researcher, who has analyzed specific requirements. Using the table is possible to find the theoretical value of t - statistics.

If the calculated value is more than theoretical t-statistics ($t_U > t$), then the null hypothesis is rejected and with the selected probability it can be argued that the studied characteristic is *statistically significant*.

Which comes first, theoretical hypothesizing or empirical research? There is a bit like asking which came first, the chicken or the egg. In practice, theorizing and experimentation feed on each other, and questions of this type cannot be answered. For this reason, we will approach the topic of hypothesis testing from both directions. On the one hand, we may suppose that the theory has come first and that the purpose of the experiment is to evaluate its plausibility. This will lead to the execution of significance tests. Alternatively, we may perform the experiment first and then consider what theoretical hypotheses would be consistent with the results. This will lead to the construction of confidence intervals.[1]

Usually quality test models test the significance of the model parameters and the determination coefficient (for regression model). Confidence interval can constructed for the investigated variables using the tabular value of t-statistics. *The confidence interval* is the interval we can expect the actual values of the studied variable with a certain probability.

The confidence interval will be determined by the formula:

$$U - t \cdot s_u \le M(U) \ \ \le U + t \cdot s_u \text{(2.11)}$$

2.5 QUALITY OF SIMPLE LINEAR REGRESSION MODEL

Practical use of econometric models has a great importance, especially their adequacy, i.e. according to the real process and the statistical data used in model construction. The quality analysis (verification of model) includes statistical and substantive part. Tests of statistical quality econometric models usually consists of the following steps:

- Check the overall quality of the regression equation.
- Check the statistical significance of the coefficients of the regression equation.
- Checking the preciseness of the model.
- Check the properties of the data, the execution of which was assumed in the evaluation of the equation (for example, the conditions of the Gauss - Markov).

Under the substantive component of the quality analysis refers to the examination of the economic meaning of the resulting model and its coefficients.

2.6 ASSESSMENT OF THE STATISTICAL SIGNIFICANCE OF THE MODEL PARAMETERS

The equation of the model defined by the selected source data has the form: $\hat{Y} = a + b \cdot X$ (simple linear regression). The model parameters are random variables(a and b, calculated according to the sample). Their mathematical expectations is equal to α and β when performing assumptions about the variance ε_i. The estimation of to α and β is the more, the less they spread around , i.e. less than their variance.

$$S_b^2 = \frac{S^2 e}{\sum_{i=1}^{n} (x_i - \overline{X})^2} ,$$

When checking the quality of the model it is necessary to check existence of a linear relationship between X and Y, i.e. to check the statistical significance of the parameter . This analysis should be done according to the scheme of statistical testing of hypotheses, as already mentioned. Formulated two hypotheses:

H0: b=0

H1: b≠0

It is calculated t-statistics $t_b = \frac{|b|}{S_b}$.

You can prove that it is possible to calculat S_b^2 by the formula:

$$S_b^2 = \frac{S^2 e}{\sum_{i=1}^{n} (x_i - \overline{X})^2}, (2.12)$$

$S_e^{\;2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ is the estimated residual dispersion (estimation of the dispersion of the errors),

S_b - the standard deviation of the random variable b.

The value of b is a measure of the slope of the regression line. Obviously, the larger the variance of the Y values around the regression line (more S_e^2) lead larger (average) error in the determination of the tilt of the regression line. If there is no variation at all (the ε_i=0 and hence S_e^2=0), then the linear function is unambiguous and there is no error in the determination of the parameters.

The denominator value S_b depends on the range of variation of the variable X. The wider value leads bigger $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and smaller error in the estimation of the tilt line. In addition, increasing the number of observations (ceteris paribus) enlarges $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ and, consequently, reduces the magnitude of the error.

If t_b (the calculated value of t-statistics) will be more theoretical ($t_{b>}$t), then the null hypothesis is rejected and the coefficient b is recognized as statistically significant with the selected confidence level (α). In this case, it is possible to construct a confidence interval the β coefficient:

$$b - t \cdot S_b \leq \beta \leq b + t \cdot S_b \quad (2.13)$$

If the null hypothesis is accepted (b=0), it indicates the absence of a relationship between the explained and factor variables.

On a similar scheme tests the hypothesis of statistical significance of the coefficient a:

$$D(a) = S_a^2 = \cdot \frac{S_e^2 \cdot \sum_{i=1}^{n} X_i^2}{n \cdot \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = S_b^2 \cdot \frac{\sum_{i=1}^{n} X_i^2}{n} \quad (2.14)$$

S_a and S_b is the standard dispersion of random variables a and b.

The dispersion of the free member of the equation is proportional S_b^2, so for it fair so as fair early made explanations about the influence of the scatter Y_i around the regression line and the scatter X_i of the standard error.

The more changes the slope of a line drawn through a given point $\left(\overline{X}, \overline{Y}\right)$, the greater the scatter free member, describing the point of intersection of this line with the y-axis.

In addition, dispersion and standard error of the free member is greater when greater the average of value $\overline{X^2}$. A small change in the tilt of regression line can cause meaningful change in free member, when module of values X is large. In this case the distance become large from point observations to the y-axis.

For statistically significant parameters of the model it is possible to construct confidence intervals, the value of which forms an interval estimate of the corresponding true model parameters.

2.7 THE COEFFICIENT OF DETERMINATION. THE STATISTICAL SIGNIFICANCE OF THE COEFFICIENT OF DETERMINATION.

It is desirable to have the indicator measuring how the regression function is determined by the factor (explaining) variables X and a stochastic perturbation ε. The characteristic would evaluate the adequacy of the model or the degree of coherence between the estimated and actual values of Y.

At first glance it seems that the qualitative criteria of assessment could serve a sum of squares of deviations of the actual values from the dependent variable calculated by the estimated equation values $\hat{Y}_i$. However, this value depends on the units of the dependent variable Y and the number of observations in the sample, so it isn’t good for evaluation.

The variance of the random variable Y in the sample can be measured using the variation (dispersion):

For analysis of the regression models will carry out the decomposition of this value into components (rule decomposition of the variance).

$$D(Y) = 1/n \cdot \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

It is obvious that:

$$Y_i - \overline{Y} = (Y - \hat{Y}_i) + (\hat{Y}_i - \overline{Y}),$$

$Y_i - \hat{Y}_i = e_i$ (graphical illustration is shown in figure 2.2).

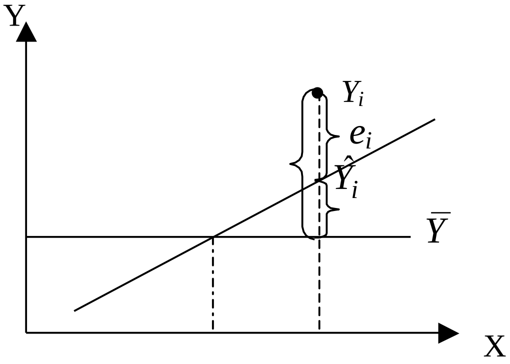

Fig.2.1. Decomposition of deviations from the sample mean

So as $Y_i = \hat{Y}_i + e_i$, $D(Y) = D(\hat{Y} + e) = D(\hat{Y}) + D(e) + 2\operatorname{cov}(\hat{Y}, e)$

It is easy to check that $\operatorname{cov}(\hat{Y}, e) = 0$. Then we have the following equality, called the decomposition rule variations:

$D(Y) = D(\hat{Y}) + D(e)$, (2.15)

Here you can write a ratio:

$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. (2.16)

The dispersion of actual values Y_i around the average measured full sum of squares TSS = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = nD(Y)$ is the total deviation (variation).

The amount of the ESS .= $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = nD(\hat{Y})$ defines the dispersion of theoretical values around the middle and is called factorial deviation (variation). It is a measure of dispersion calculated (theoretical) values determined are included in the equation independent variables $\hat{Y}_i$, so this deviation is called explained.

RSS $= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} e_i^2 = nD(e)$ residual variance. This deviation cannot be explained by the correlation dependence between Y and X, hence its name: "the unexplained", or residual variance. It measures the part of the scattering that occurs due to various random factors. Therefore, the closer RSS to zero, the less the actual Y values deviate from calculated according to equation model values $\hat{Y}$.

The relation (2.15) we write as

TSS=ESS+RSS.

Divide this value by TSS

$$\frac{ESS}{TSS} + \frac{RSS}{TSS} = 1 \qquad (2.17)$$

Value $R^2 = \frac{ESS}{TSS}$ is called the coefficient of determination (a measure of certainty). Otherwise you can also write: $R^2 = 1 - \frac{RSS}{TSS}$

R^2 shows the proportion of common variation in the analyzed dependent variable caused by a change of the factor variables. For the case of simple linear regression R^2 is equal to the square of the correlation coefficient of variables Y and X(R^2_{YX}).

Equation (2.17) shows that the lower RSS, the closer R^2 to 1 and the better the model. In the General case, the numerical value of the determination coefficient is between zero and one: $0 \le R^2 \le 1$.

If R^2 = 1, the empirical values of Y are located on the regression line. If the coefficient of determination is equal to zero, then between X and Y there is no correlation and the regression line parallel to the X-axis. Thus, if there is a statistically significant linear relationship between the values X and Y, the coefficient of determination should be close to 1.

The aim of regression analysis is to explain the behavior of the dependent variable Y. In any given sample, Y is relatively low in some observations and relatively high in others. We want to know why. The variations in Y in any sample can be summarized by the sample variance. We should like to be able to account for the size of this variance.[1]

However, you should not place too much emphasis on high value R^2, as the coefficient of determination can be close to 1 due to the fact that both values studed { and Y have a pronounced time trend, not associated with their causal dependence. In the economy generally, such trend has aggregated indicators (GNP, GDP, income, etc). Therefore, the construction and evaluation of models for time series of volume indices value may be very close to 1, which does not necessarily indicate the presence of significant linear coddelation between the studied indicators.

If the regression equation is based on cross-data, the coefficient of determination can be low even when a quality of model is satisfactory due to the high variations between the individual elements, usually R^2 does not exceed 0,7 in such situation. The same holds generally for regression time series, if they haven't a significant trend. In macroeconomics examples of such dependencies are due to the relative, specific, growth indicators, for example: the dependence of the inflation rate from unemployment; accumulation rates on the value of the interest rate and other.

It is impossible to specify the exact boundary of acceptability for all case immediately. Researchers can be guided by the assessment of the link shown in the following table.

Table 2.1 The assessment of the link based on R^2

Values R^2	[0,1-0,3)	[0,3-0,5)	[0,5-0,7)	[0,7-0,9)	[0.9-0,99]
Degree relationships	weak	moderate	noticeable	high	very high

When researcher gets the value of $R^2 < 0,3$, he must re-specification of the model.

In other cases, it is necessary to consider whether included in the model variables absolute or relative, do they have a time trend, sample size, etc.

Next step is assessment of compliance of the received data about the coefficient to determination coefficient in the General totality. It is advisable to check R^2 for statistical significance. This process can be implemented by evaluating the statistical significance of the linear correlation coefficient between X and Y according to the scheme of t-testing of hypotheses especially for simple linear models. In this case, the standard error is calculated using the following formula:

$$S_R = \sqrt{\frac{1-R^2}{n-2}}$$

2.8 EVALUATION OF THE ACCURACY OF THE MODEL

The actual values of the explained variable differ from the theoretical (calculated according to model equation) in the value of $e_i = Y_i - \hat{Y}_i$. It is possible to calculate the value in each observation and name the approximation error. Deviations $e_i = Y_i - \hat{Y}_i$ are absolute approximation error, but they are not comparable among themselves (in different models because of mesurments). So, if one observation turned out to be error 5, and in another 10, this does not mean that in this case the first or second model gives the worst result. Therefore, in order scores were comparable, considering the relationship of variance to actual values (in percent). There $e_i = Y_i - \hat{Y}_i$ can be both positive and negative value, when determining the approximation errors for each observation variances are taken module.

Value $\delta_i = \frac{|Y_i - \hat{Y}_i|}{Y_i} \cdot 100$ can be considered as a relative approximation error in the i-th observation. It is useful to determine the average relative approximation error in order to have a General opinion about the accuracy of the model:

$$\delta = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|Y_i - \hat{Y}_i|}{Y_i} \cdot 100 = \frac{1}{n} \cdot \sum_{i=1}^{n} \frac{|e_i|}{Y_i} \cdot 100. \ (2.18)$$

It is possible to define approximation errors using the standard deviation:

$$\delta = \frac{1}{\overline{Y}} \cdot \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n}} \cdot 100. \ (2.20)$$

The value of the approximation errors up to 15-20% allows to conclude about a sufficient the accuracy of the model.

Model improvements should not be measured in terms of accuracy gains. It may be going too far to say that accuracy is irrelevant, but caution is advised when using accuracy in the evaluation of predictive models.

2.9 POINT AND INTERVAL ESTIMATES OF THE DEPENDENT VARIABLE

One of the main tasks of econometric modeling is to predict values of the dependent variable for specific values of the explanatory variables.

Let the equation of the model defined by the sampling data has the form:

$$\hat{Y} = a + b \cdot X \ .. \ (2.21)$$

Parameters a and b contain random errors. In the dependent variable $\hat{Y}(X_0)$ (was found using equation model at some point) also contains random errors and, therefore, defines a conditional mean value of Y at the point X_0 (the point estimate). Let denote it Y_X.

It can be shown that the variance of this quantity is calculated by the formula:

$$S_{Y_X}^2 = DY(X) = S_e^2 \cdot \left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] \ (2.22)$$

Provided the requirements of normal distribution of the residuals ε random variable Y_X has a normal distribution, and statistics $t = \frac{Y_X - M_X(Y)}{S_{Y_X}}$ has a t-distribution with number degrees of freedom (n-2). Then for the conditional mathematical expectation $M_X(Y)$ can be found confidence interval:

$$Y_X - t \cdot S_{Y_X} \le M_X(Y) \le Y_X + t \cdot S_{Y_X}\ , (2.23)$$

$S_{Y_X} = \sqrt{S^2_{Y_X}}$ is the standard error of the conditional average of the dependent variable.

The formula (2.22) and (2.23) shows that the width of the confidence interval depends on the value X$_0$: if $X_0 = \overline{X}$ it is minimal. If the distance average and analyzed X is growing, the average width of the confidence interval increases (figure 2.2).

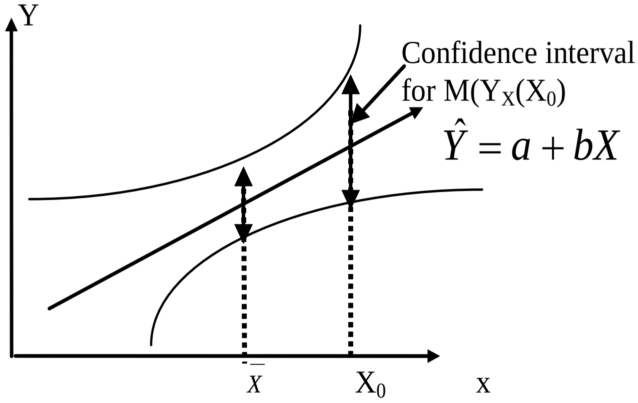

Fig. 2.2 The confidence interval for explained variable

Figure 2.2 shows confidence region, which determines the location of the model the regression line, i.e. the conditional mathematical expectation, but doesn't give separate values of the dependent variable, which may vary around the average.

Sometimes we can be more interested in finding confidence intervals for for some individual values of Y, that we associate with X$_0$, than the average values of Y.

In a slightly different form, this problem can be formulated as: when receiving a new pair of observations (X$_{0,}$ Y$_0$) to find out whether it is still based on equation formulated, i.e. equals Y$_0$ to $\hat{Y}_0$ (value obtained by substitution in the equation of the model X$_0$).

Let's consider the value $z = Y_0 - \hat{Y}_0$. We formulate two hypotheses:

H$_0$: z = 0;

H$_1$: z $\neq$ 0

It is possible to calculate that the estimated variance of the magnitude of z is calculated by the formula:

$$S_z^2 = S_e^2 \cdot \left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2} \right] \quad (2.24)$$

Because the variable $z = Y_0 - \hat{Y}_0$ is a linear combination of normally distributed variables, it also has a normal distribution. Therefore, the value $t = \dfrac{z}{S_z}$ has a t-distribution with (n-2) degrees of freedom.

If the calculated value of t-statistic is more than the tabblated, the null hypothesis is rejected, i.e. with the selected level of confidence we can assert that the value Y_0 is statistically significantly (significantly) different from the values $\hat{Y}_0$ found using equation model and the considered pair (X_0, Y_0) does not correspond to the relationship

For individual values of the variable Y can be constructed confidence interval:

$\hat{Y}_0 - t \quad \cdot s_z \le Y_0 \le \hat{Y}_0 + t \quad \cdot s_z$. (2.25)

Obviously, this interval is wider than for conventional medium (X_0) with the same level of trust and includes the confidence interval for the conditional mean.

2.10 THE USE OF ECONOMETRIC MODELS FOR FORECASTING

The forecasting process, based on an econometric model, breaks down into the following stages:

- selection and building of model;
- evaluation of the constructed model;
- forecast (point and interval).

To obtain point forecasts put the analyzed value $\overline{X}_0$ in the equation model and find $Y(\overline{X}_0)$. This is the point forecast.

However, the probability of Y is the found point $\hat{Y}_0$ is almost zero, therefore there is a need for projections of a "plug" through confidence intervals - *interval forecast*. Interval forecast can be constructed for average and individual values.

The analyzed value $\overline{X}_0$ may lie both within sample and out of it. At the same time, if $\overline{X}_0$ is outside the sample and very different from the average level, width of the confidence interval increases significantly. This demonstrates the vagueness of the forecast.

Data (based on the forecast) should be critically conceptualized with a meaningful point of view.

2.11 THE EXAMPLE OF CONSTRUCTION AND QUALITY ASSURANCE SIMPLE REGRESSION MODEL

It is known a statistical data (table. 2.3) about the change of the output Y in leading to changes in costs of labor X.

Assignments:

1. To build a model of the relationship between these indicators (X and Y).
2. To check the quality of the constructed model

A. To check the overall quality built model

B. To assess the statistical significance of the model parameters.

3. To estimate the parameters of the true model.
4. To assess the accuracy of the model.
5. To construct point and interval forecast for the cost of labor in the amount of

39 thousand rubles

Table 2.3 Output and labor coast.

Output	20	22	25	28	30	34	35	44	50	56
Costs of labor	8	10	18	22	34	46	50	54	58	60

1. Model building

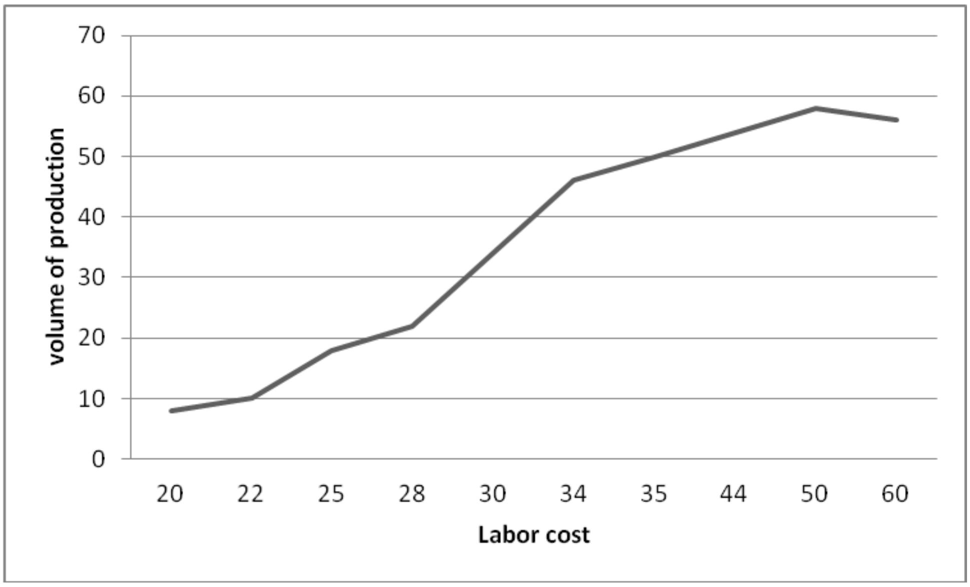

Figure 2.3 Flowchart the change of the output Y in leading to changes in costs of labor X.

Based on the analysis of the flowchart (Fig. 2.3) we can suppose that between the studied indicators there is a linear relationship: $Y = \alpha + \beta \cdot X + \varepsilon$. Estimate the parameters of this model based on the method of least squares. The estimated equation models: $Y = \alpha + \beta \cdot X + \varepsilon$

Make the table to calculate the parameters and characteristics of the model.

Table 2.4 Supporting calculations

N	X	Y	X^2	XY	$\hat{Y}$	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{Y})^2$	e^2	$(X-\bar{X})^2$	e	$\frac{\lvert e_i \rvert}{y_i}$
1	8	20	64	160	19	219	256	1	762	1	0,1
2	10	22	100	220	20	164	219	4	655	2	0,1
3	18	25	324	450	25	96	100	0	310	0	0,0
4	22	28	484	616	27	46	58	1	185	1	0,0
5	34	30	1156	1020	34	23	0	19	3	-4	0,1
6	46	34	2116	1564	42	1	46	58	108	-8	0,2
7	50	35	2500	1750	44	0	85	81	207	-9	0,3
8	54	44	2916	2376	46	85	135	6	339	-2	0,1
9	58	50	3364	2900	49	231	196	1	502	1	0,0

10	56	60	3136	3360	48	635	164	154	416	12	0,2
∑	356	348	16160	14416	354	TSS= 1500	ESS= 1258	RSS= 325	3486		1,1
ср.	36	35	1616	1442	35						

We write the system of normal equations and find the solution.

$$\begin{cases} a + 36 \cdot b = 35 \\ 36 \cdot a + 1616 \cdot b = 1442 \end{cases}$$

$$b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{1442 - 35 * 36}{1616 - 36^2} \approx 0{,}6\,;$$

$$a = 35 - 0{,}6 * 36 \approx 14$$

Received the following equation model: $\hat{Y} = 14$+0,6*X.

2. Quality equation regression models

A. To evaluate the overall quality of the model (adequacy).

Calculate the coefficient of determination.

$R^2 = \dfrac{ESS}{TSS} = \dfrac{1258}{1500} \approx 0{,}8$ indicates a high degree of correlation between Y and X.

Will check on the statistical significance of the correlation coefficient R= $R = \sqrt{R^2}$ . .

Find the theoretical value of t-test at 95% confidence level and the number of degrees of freedom $n-2=10-2=8$ from table of t - distribution (see Appendix) : =2,306.

Find the estimated value of a statistical t-test:

$t_R = \dfrac{|R|}{S_R}$, $S_R = \sqrt{\dfrac{1-R^2}{n-2}} = \sqrt{\dfrac{1-0{,}08}{8}} = 0{,}34$ then $t_R = \dfrac{\sqrt{0{,}8}}{0{,}34} \approx 2{,}63$, more than the tabulated value of t-statistic.

Therefore, the linear correlation coefficient and determination coefficient of the linear model are statistically significant. Since the coefficient of determination is characterized by high connectivity factor and dependent variables, the model can be considered as adequate.

B. Checking the statistical significance of the model parameters.

Since the model building is based on the sample data, it is necessary to test the statistical significance of the model parameters a and b.

For parameter b: $S_b^2 = \dfrac{S_e^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{40{,}6}{3486} \approx 0{,}01;$

S_e^2 is the estimated residual variance: $S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{325}{8} \approx 40{,}6.$ Then $t_b = \frac{0{,}6}{\sqrt{0{,}01}} = 6$

For the parameter a: $S_a^2 = S_b^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n} = 0{,}01 \cdot \frac{1616}{10} \approx 1{,}62$

$t_a = \frac{14}{\sqrt{1{,}62}} \approx 11$

Find the theoretical value of a statistical t-test at 95% confidence level and the number of degrees of freedom $n-2=10-2=8$. Using table of t-statistics (see Appendix): t=2,306.

Since $t_b > t$ and $t_a > t$, both the equations of the model are recognized as statistically significant with a probability of 95%.

Note: the Statistical significance of the parameter b confirms the link between output and the cost of fixed assets.

3. Estimate the parameters of the true model

To estimate the parameters of the true model, α and β, use confidence interval for statistically significant estimated parameters of the model. In the model both parameters are statistically significant, therefore, it is possible to construct confidence intervals for α and β.

The parameter α:

$a - t_T \cdot S_a \le \alpha \le a + t_T \cdot S_a$

$14 - 2{,}306 \cdot \sqrt{1{,}62} \le \alpha \le 14 - 2{,}306 \cdot \sqrt{1{,}62}$

Therefore, the true model parameter α ϵ [11,1:16,9]

The parameter β:

$b - t \cdot S_b \le \beta \le b + t \cdot S_b$

$0{,}6 - 2{,}306 \cdot \sqrt{0{,}01} \le \beta \le 0{,}6 + 2{,}306 \cdot \sqrt{0{,}01}$

Therefore, the true model parameter β ϵ [0,4:0,8]

4. The accuracy of the model

The accuracy of the model is determined on the basis of the average relative approximation errors: $\delta = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| \frac{e_i}{y_i} \right| = 1{,}1\% < 10\%$ - the model is accurate.

Conclusion: the analysis of the model shows that it is adequate and accurate model.

5. The prediction based on the model

To locate a point forecast, we will substitute X=40 in the equation model Y(40)=14+0,6·40≈38.

When labor costs is 40 thousand rubles output will account for 38 pieces.

Find the forecast interval for average values of Y for a given value of X (the conditional mean). To do this, let us calculate the dispersion:

$$S^2_{Y(X_0)} = S^2_e(\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum\limits_{i=1}^{n}(X_i - \overline{X})^2}) = 40,6 \cdot (\frac{1}{10} + \frac{(40-36)^2}{3486}) = 4,2$$

The tabular value of t-statistics was found previously (t=2,306). Then the confidence interval (95%condence level) for the average value of Y when X_0=40:

$$38 - 2,306 \cdot \sqrt{4,2} \le M(Y_X(X_0 = 20) \le 38 - 2,306 \cdot \sqrt{4,2}$$

Consequently, the average volume of production at a cost of fixed assets 40 units with a probability of 95% will be in the interval:

$$33,2 \le M(Y_X(X_0 = 20)) \le 42,7$$

Appendix

Statistic tables.

Table of values of F- distribution (Fisher).

The significance level α=0,05.

k1 k_2	1	2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,5	19,00	19,16	19,25	19:30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38

Table of values of $\chi 2$ (Chi-square) distribution.

df\area	**1**	**0,99**	**0,98**	**0,95**	**0,9**	**0,75**	**0,5**	**0,25**	**0,1**	**0,05**	**0,03**	**0,01**	**0,01**
1	0,00	0,00	0,00	0,00	0,02	0,10	0,45	1,32	2,71	3,84	5,02	6,63	7,88
2	0,01	0,02	0,05	0,10	0,21	0,58	1,39	2,77	4,61	5,99	7,38	9,21	10,60
3	0,07	0,11	0,22	0,35	0,58	1,21	2,37	4,11	6,25	7,81	9,35	11,34	12,84
4	0,21	0,30	0,48	0,71	1,06	1,92	3,36	5,39	7,78	9,49	11,14	13,28	14,86
5	0,41	0,55	0,83	1,15	1,61	2,67	4,35	6,63	9,24	11,07	12,83	15,09	16,75
6	0,68	0,87	1,24	1,64	2,20	3,45	5,35	7,84	10,64	12,59	14,45	16,81	18,55
7	0,99	1,24	1,69	2,17	2,83	4,25	6,35	9,04	12,02	14,07	16,01	18,48	20,28

8	1,34	1,65	2,18	2,73	3,49	5,07	7,34	10,22	13,36	15,51	17,53	20,09	21,95
9	1,73	2,09	2,70	3,33	4,17	5,90	8,34	11,39	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	6,74	9,34	12,55	15,99	18,31	20,48	23,21	25,19
11	2,60	3,05	3,82	4,57	5,58	7,58	10,34	13,70	17,28	19,68	21,92	24,72	26,76
12	3,07	3,57	4,40	5,23	6,30	8,44	11,34	14,85	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	9,30	12,34	15,98	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	10,17	13,34	17,12	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	11,04	14,34	18,25	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	11,91	15,34	19,37	23,54	26,30	28,85	32,00	34,27
17	5,70	6,41	7,56	8,67	10,09	12,79	16,34	20,49	24,77	27,59	30,19	33,41	35,72
18	6,26	7,01	8,23	9,39	10,86	13,68	17,34	21,60	25,99	28,87	31,53	34,81	37,16
19	6,84	7,63	8,91	10,12	11,65	14,56	18,34	22,72	27,20	30,14	32,85	36,19	38,58
20	7,43	8,26	9,59	10,85	12,44	15,45	19,34	23,83	28,41	31,41	34,17	37,57	40,00

Table of values of T-statistic (Student).

Number of freedom levels		**α**					
		0,2	**0,1**	**0,05**	**0,02**	**0,01**	**0,002**
1		3,078	6,314	12,706	31,821	63,657	636,619
2		1,886	2,920	4,303	6,965	9,925	31,599
3		1,638	2,353	3,182	4,541	5,841	12,924
4		1,533	2,132	2,776	3,747	4,604	8,610
5		1,476	2,015	2,571	3,365	4,032	6,869
6		1,440	1,943	2,447	3,143	3,707	5,959
7		1,415	1,895	2,365	2,998	3,499	5,408
8		1,397	1,860	2,306	2,896	3,355	5,041
9		1,383	1,833	2,262	2,821	3,250	4,781
10		1,372	1,812	2,228	2,764	3,169	4,587
11		1,363	1,796	2,201	2,718	3,106	4,437
12		1,356	1,782	2,179	2,681	3,055	4,318
13		1,350	1,771	2,160	2,650	3,012	4,221
14		1,345	1,761	2,145	2,624	2,977	4,141
15		1,341	1,753	2,131	2,602	2,947	4,073

16		1,337	1,746	2,120	2,583	2,921	4,015
17		1,333	1,740	2,110	2,567	2,898	3,965
18		1,330	1,734	2,101	2,552	2,878	3,922
19		1,328	1,729	2,093	2,539	2,861	3,883
20		1,325 **0,100**	1,725 **0,05**	2,086 **0,025**	2,528 **0,01**	2,845 **0,005**	3,850 **0,001**

Table of values of d– statistics Darbin-Watson (the values of d1 and d2 at 5% significance level)

n	*k*=1		*k*=2		*k*=3		*k*=4		*k*=5		*k*=6		*k*=7	
	*d*1	*d*2	*d*1	*d*2	*d*1	*d*2	*d*1	*d*2	*d*1	*d*2	*d*1	*d*2	*d*1	*d*2
6	0,61	1,40												
7	0,70	1,36	0,47	1,90										
8	0,76	1,33	0,56	1,78	0,37	2,29								
9	0,82	1,32	0,63	1,70	0,46	2,13	0,30	2,59						
10	0,88	1,32	0,70	1,64	1,53	2,02	0,38	2,41	0,24	2,82				
11	0,93	1,32	0,76	1,60	0,60	1,93	0,44	2,28	0,32	2,65	0,20	3,01		
12	0,97	1,33	0,81	1,58	0,66	1,86	0,51	2,18	0,38	2,51	0,27	2,83	0,17	3,15
13	1,01	1,34	0,86	1,56	0,72	1,82	0,57	2,09	0,45	2,39	0,33	2,69	0,23	2,99
14	1,05	1,35	0,91	1,55	0,77	1,78	0,63	2,03	0,51	2,30	0,39	2,57	0,29	2,85
15	1,08	1,36	0,95	1,54	0,81	1,75	0,69	1,98	0,56	2,22	0,45	2,47	0,34	2,73

4. Table of values of RS– statistics (at 5% significance level)

n	Lower level(a)	Upper level(b)
8	2,5	3,4
9	2,58	3,54
10	2,67	3,68
12	2,8	3,9
14	2,92	4,09
16	3,01	4,24
18	3,1	4,37
20	3,18	4,49
25	3,18	4,71
30	3,47	4,89

Литература

1. Dougherty C. Introduction to econometrics. Fourth Edition. Oxford University Press, 2011.
2. Green W. Econometric Analysis. Seventh Edition. N.Y. University, 2011.
3. Gujarati D. Basic Econometrics. Third edition. McGraw-Hill. 1995.
4. Johnston J., Dinardo J. Econometric Methods. Fourth Edition. MacGraw-Hill, 1997.
5. Luetkepohl H. Applied Time Series Econometrics. Cambridge University Press, 2009.
6. Wooldridge J. Introductory Econometrics: A Modern Approach. Third edition. Thomson South-Western, 2005

Марина Владимировна **Кемаева**

ЭКОНОМЕТРИКА

Учебно-методическое пособие

Федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Нижегородский государственный
университет им. Н.И. Лобачевского».
603950, Нижний Новгород, пр. Гагарина, 23.