MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE
WEST UKRAINIAN NATIONAL UNIVERSITY

BASIC CONSPECTUS OF LECTURES
in discipline
«ECONOMETRICS»

Ternopil – 2022

UDK 519.2

Reviewers:

S.V. Martyniuk – PhD, Candidate of physical and mathematical sciences, associate professor of the Department of Mathematics and Teaching Methods at Volodymyr Hnatyuk Ternopil National Pedagogical University.

O.S. Bashutska – PhD, Candidate of Economics, associate Professor of the Department of Economic Cybernetics and Informatics of the West Ukrainian National University;

approved at the meeting of the Department of Applied Mathematics, protocol № 1 of 26.08.2022.

Basic conspectus of lectures in discipline «Econometrics» is considered basic theoretical knowledge to study the discipline «Econometrics».

Plaskon S.A., Dzyubanovska N.V. Basic conspectus of lectures in discipline «Econometrics». - Ternopil: WUNU, 2022.- 45 p.

UDK 519.2

Responsible for release: O.M. Martyniuk, PhD, Candidate of physical and mathematical sciences, associate professor of the Department of Applied Mathematics, Head of the Department of Applied Mathematics WUNU

«Econometrics» is one of the basic disciplines, which gives students the opportunity to learn the calculation method, theoretical analysis and practical application of correlation-regression models, which is necessary for researching the functioning and development of economic systems with the possibility of forecasting the studied factors in various areas of the economy.

The goal of the discipline "Econometrics" is to master a set of mathematical methods used for quantitative assessment of economic phenomena and processes; learning econometric modelling, construction economic-mathematical models, the parameters of which are estimated by means of mathematical statistics; learning empirical derivation of laws; preparation for applied research in the field of economics; mastering the mathematical apparatus that helps to analyze, model and solve applied economic problems; development of students' logical and algorithmic thinking; teaching them methods of solving mathematically formalized problems; instilling in them the skills of independent study of scientific and reference literature.

Topic. Econometrics and econometric modeling:
basic concepts and definitions

1.1. Econometrics and its connection with mathematical and statistical methods

Market relations open a wide space for the use of researched econometric methods, which make it possible not only to carry out quantitative calculations, but also to choose optimal predictive scenarios of actions. Even if these scenarios are not easy to apply, and the constraints of the model are not strictly fulfilled, they serve as a reasonable starting point to guide rational decision-making.

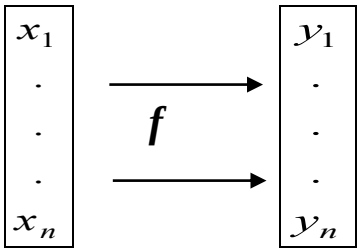
Econometrics is an independent economic-mathematical discipline that unites a set of theoretical results, methods, methods and models designed to give a specific quantitative expression to general (qualitative) regularities on the basis of economic theory, statistics of mathematical economics and mathematical-statistical tools and predict the results of the development of complex economic processes. Thus, the essence of econometrics lies in the synthesis of three components: economics, economic statistics and mathematics. Considering economic theory within the framework of econometrics, we will be interested not only in the identification of objectively existing economic laws and relationships between economic indicators, but also in approaches to formalization, including methods of specification of relevant models, taking into account the problem of their identification. When using economic statistics as a component of econometrics, we will be interested only in the aspect related to the information provision of the analyzing econometric model. The mathematical and statistical tools of econometrics do not, of course, mean mathematical statistics in the traditional sense, but only its separate sections (regression and dispersion analysis, time series, analysis of the system of simultaneous equations, etc.)

It follows from the definition of econometrics that its origin and main purpose are socio-economic applications and a model description of the existing quantitative

relationships and interdependencies between the analyzing indicators of economic processes.

Finding dependencies and interrelationships between objectively existing phenomena and processes of the economy is of great importance, which in turn makes it possible to better understand the complex mechanism of cause-and-effect relationships between them. Currently, objectively existing dependencies and relationships between economic phenomena are mostly described verbally (qualitatively). Along with this, the quantitative measurement of cause-and-effect relationships inherent in random processes is more important. To study the intensity, type and form of causal influences in the middle of stochastic processes, methods of multivariate statistical analysis are used, among which a special role is assigned to correlation and regression analysis. Taking into account the nature of phenomena inherent in economic processes, the mathematical apparatus of correlation-regression analysis allows you to create stochastic models and show their superiority in this field in comparison with deterministic models. If relative to the quantities analyzed in deterministic models, an assumption is made regarding their stability, and random deviations are not taken into account, then in stochastic models, the random nature of indicators is taken into account, and they are estimated by probability measures.

Deterministic models describe regularities that appear individually in each individual element of the population. The relationship between cause and effect in laws of this type can be expressed quite precisely in the form of specific mathematical formulas, systems of equations, since certain quantitative values of influential factors (arguments) always correspond to certain values of the resulting indicator (function). Such a connection is called functional. With a functional dependence between the variables X and Y, every possible value of Xi is unique corresponds to a certain value of Yi. The functional dependence can be written in the form: $y=f(x)$ or represented schematically:



Regularities that appear in mass cases, only with a large number of observations, are called statistical. Statistical regularities are causally conditioned, the existing set of causes are interconnected and act in different directions. In such conditions, it is difficult to detect a quantitative relationship between cause and effect. The analytical expression of statistical regularities is determined by the methods of mathematical statistics. A cause-and-effect relationship due to the simultaneous action of many causes is clearly manifested only in the mass of cases, it is called correlational or stochastic, and it is characteristic of statistical regularities. In economics, one often has to deal with many phenomena that are probabilistic in nature. With stochastic dependence, for given values of the independent variable X, it is possible to specify a number of values of the dependent variable Y, which are randomly scattered in a certain neighborhood. Each fixed value of the argument corresponds to a certain statistical distribution of function values. This is due to the fact that, in addition to the selected variable X, the dependent variable is affected by a number of uncontrollable or unaccounted for factors, as well as measurement errors. Since the value of the dependent variable is subject to random dispersion, it cannot be predicted with sufficient accuracy, but only specified with a certain probability. At the same time, economic processes are characterized by random deviations and interrelationships in time, therefore, for the analysis of a complex of cause-and-effect relationships, stochastic models. With the help of these models, it is possible to construct predictive estimates of the parameters of the functioning of economic systems in dynamics.

The most famous methods of building stochastic models are correlation and regression analysis. In applied economic research, they are precisely the tool that can identify and evaluate a complex set of relationships and consequences.

The results of the analysis of the conducted studies showed that the modeling of economic processes with the help of one regression equation is quite primitive and insufficient due to the existing set of cause-and-effect relationships. For their more adequate display, it is necessary to use a system of simultaneous equations.

Let's consider the classification of problems that are solved by the mathematical apparatus of econometrics according to the following characteristics: final applied goals, level of hierarchy and profile of the analyzed economic system.

According to the final applied goals, we will highlight two main tasks:

1) forecast of economic and socio-economic indicators that characterize the state and development of the analyzing system;

2) simulation of various possible scenarios of socio-economic development of the analyzing system, when relationships between the characteristics of production, consumption, social and financial policy, etc. are statistically revealed. are used to study how possible changes in certain production or distribution control parameters will affect the values of the output characteristics.

Macro-level, meso-level, and micro-level tasks are distinguished according to the level of the analysis economic system hierarchy.

In some cases, the profile of econometric modeling should be defined: the research may be focused on market problems, investment, financial, tax or social policy, pricing, demand and consumption, or on a certain set of problems.

Consider a scheme that more fully defines econometrics and shows its relationship with mathematical, statistical and economic disciplines (Fig. 1.1).

We can single out five main tasks that econometrics solves.

First, the model must be specified, that is, all functional connections must be included in it in an explicit form. Econometrics can reach this by going from simple to complex: starting with the simplest functions, introducing and testing various hypotheses, and gradually complicating the nature of functional relationships based on real data.

Secondly, the task of econometrics is the choice of definition and measurement of the variables included in the model.

Thirdly, it is necessary to estimate all unknown parameters of the models and calculate confidence intervals (intervals into which the calculated value will fall with a given degree of probability).

Fourth, it is necessary to evaluate the quality of the built models using various tests and criteria. This helps to finally decide whether the initially selected model and some theoretical assumptions should be changed.

If such a variable is necessary, then new calculations and new testing must be carried out.

Fifth, having a final model, it is necessary to conduct a deep analysis of the results that are planned to be used in practice for decision-making.

1.2. Econometric model and stages of econometric modeling

The main fundamental idea that we encounter when studying economics is the idea of the relationship between economic parameters. The market-forming demand for a product is considered as a function of its price; costs associated with the production of any product are assumed to depend on the volume of production; the amount of tax revenues to the budget can be a function of the tax rate, etc. All these are examples of the relationship between two variables, one of which (demand for a product, production costs, consumer spending, the amount of tax revenues) plays the role of a performance indicator, and the others are interpreted as explanatory factors (factors-arguments) with the help of which the performance indicator is formed . However, for greater reality, several explanatory variables and a residual random component should be introduced into each such ratio, which reflects the impact on the performance indicator of all unaccounted factors. The demand for a good can be seen as a function of its price, consumer income, and the price of competing goods; consumer spending can be defined as a function of income, liquid assets, and the previous level of consumption; the amount of tax revenues to the budget will depend on the profit received and the tax rate. At the

same time, the random component taking part in each of these ratios reflects the impact on the analytical performance indicator of all unaccounted factors and determines the stochastic nature of the dependence, namely: even after fixing the values of the explanatory variables at certain levels, for example, the prices of the product itself and those competing with it, and as well as consumer income, we cannot expect that this will unequivocally determine the demand for this product. In other words, moving in our observations of demand from one temporal or spatial stage to another, we will find a random variation in the value of demand around a certain level, even while keeping the values of all argument factors unchanged. Taking into account the main tasks that econometrics solves, we will present the logical structure of the econometric modeling process in the form of such a scheme.

Econometric models are an important class of models offered by mathematics to analytics. With the help of these models, phenomena are described in which there are statistical factors that are difficult to explain in a purely deterministic approach. Typical models of this kind have a trend-cyclical component and a random component. Whether he likes it or not, the analyst cannot exclude the random component and must build his predictive conclusions, taking into account their presence.

Let's introduce a random variable u, which includes part of the variation of the performance indicator y, which is not explained by the independent variables. That is, there will be deviations between the values calculated by the model and the actual values:

$$u_i = y_i - \hat{y} \text{ або } y_i = \hat{y}_i + u_i \qquad (1.1)$$

The random variable u is called a disturbance (residual, deviation). Its values can change from one observation to another. For example, when studying the dependence of national income on capital investments, the disturbing variable would include the influence on national income of such factors as the number of people working in the field of production, labor productivity, use of fixed assets, etc., as well as other random factors.

Thus, the simplest econometric model can be presented as follows:

$$y = \hat{y} + u, \quad y = f(x) + u \text{ or } y = f(x_1,.., x_n) + u. \quad (1.2)$$

This type of notation allows us to interpret the random variable u as a variable that takes into account the wrong specification of the function, that is, the wrong choice of the form of the equation that describes the dependence.

Due to the introduction of the random variable u, the variable y also becomes random, because given the values of the explanatory variables, the variable y cannot be assigned or matched to only one specific value. If, for example, we study the dependence of the volume of tax revenues on the value of the tax rate, then by setting the value of the tax rate, you can specify the interval in which the corresponding volumes of revenues can be found.

The econometric model is based on the close unity of two important aspects - qualitative theoretical analysis of existing relationships and empirical information. The theoretical aspect is reflected in the specification of the model, which is an analytical form of the econometric model. That is, the model specification consists of a certain type of equations or functions used to build the models, has probabilistic characteristics that are characteristic of the stochastic residuals of the model.

The grouping of individual ratios in the form of a model is important when building econometric models. Any mathematical model is only a simplified formalized representation of a real object (phenomenon, process), and the skill of its construction consists in combining as much as possible the conciseness of the parameterization of the model with sufficient adequacy of the description of those aspects of the modeling of reality that are of interest. The number of connections included in the model depends on the conditions under which the model is constructed and on the completeness of the explanations we seek. For example, the traditional model of demand and supply should explain the relationship between price and volume of output, characteristic of a certain market. It contains three equations, namely: the demand equation, the supply equation, and the market reaction equation.

Let's build this econometric model. Let and - respectively, the amount of demand and supply of a certain product at a certain time on a given market; P - price per product unit. Values and depend on and these dependencies can be represented as follows: - demand function; is the supply function, where and are the perturbations for the corresponding variables. For the existence of equilibrium in the market, it is necessary that the condition be fulfilled. Therefore, the single-product market model will take the form:

$$\begin{cases} q_1 = f(P, u_1) \\ q_2 = f(P, u_2) \\ q_1 = q_2 \end{cases} \quad (1.3)$$

1.3. Causal relationships between variables

Economic processes are constantly interconnected. If we want to penetrate deeply and objectively into the essence of a phenomenon or process, it is necessary to investigate and reveal these connections. In the course of statistical analysis, the most significant relationships are quantitatively described. At the same time, in order to adequately reflect the essence of the passage of processes, special attention must be paid to the causal explanation of connections. A causal relationship is understood as such a combination of phenomena and processes of real reality, when a change in one of them is a consequence of a change in the other. It should be borne in mind that a causal relationship between individual phenomena may not always occur, but only under a certain set of conditions. These conditions must be implemented simultaneously with the action of the causes, if there are cause-and-effect relationships between the studied phenomena. A change in conditions can lead to a change in the causes of the impact, that is, to a change in the consequences. One of the important features of a causal relationship is the observance of the temporal sequence of cause and effect: the cause always precedes the effect. However, one should not identify the relationship of the active

cause with the relationship of the previous and the next, that is, not every previous event can be considered the cause of the appearance of the next.

In addition, for the correct understanding of cause-and-effect relationships, cases of coincidence and simultaneous development of phenomena are important in their evaluation. Another important feature of causal relations is its necessity, that is, under these conditions, the cause, when repeated with necessity, generates the same effect. It is also necessary to pay attention to the condition of repetition of phenomena, since only repetition provides a practical possibility of revealing connections. The difficulty is that the causes and conditions of many phenomena are only relatively stable. Over a period of time, any phenomenon undergoes continuous change, and we will never find its exact repetition. Neither the cause nor the effect is repeated exactly. A change in the cause-and-effect complex complicates the knowledge of phenomena. It is especially difficult to take into account the change of conditions in the study of causal relationships in economic processes.

Most economic phenomena are the result of many simultaneous and combined causes. When determining the connections between them, the main causes, which necessarily lead to a given consequence, must be distinguished from secondary ones. Secondary causes complicate the actions of causes that are essential in this aspect. In addition, to a certain extent, randomness is characteristic of the causal action and its determining consequence. Each process, when repeating its causal complex, is realized with a deviation from the law underlying it, due to accidents. This must be taken into account when learning the cause-and-effect complex of socio-economic phenomena. Randomness is recognized as an indispensable component of any phenomenon. Necessary and accidental connections exist objectively, regardless of human consciousness, and thus form a dialectical unity. The action of the main cause is joined by the influence of additional causes. At the same time, the directions of these influences may not coincide. In addition, random obstacles are imposed on the cause-and-effect complex. All this modifies the action of the main cause and leads to a different effect than it would be if there was

only one main cause. Unfortunately, due to insufficient cognitive resources, we are often unable to describe the entire complex set of reasons. Its description in a general form is insufficient for a complete understanding of the very essence of the phenomenon. That is why research is usually started by establishing the reasons that are essential in the given conditions and expressing the main causal relationships in quantitative form. Secondary causes, as well as the variation of causal relationships caused by a change in the conditions in which the phenomenon occurs, are considered in a single complex. This complex, as a rule, contains the influence of known significant and undisclosed causes, random obstacles, the influence of causes that cannot be solved quantitatively. The presence of an influential complex complicates economic research and makes it impossible to fully cover cause-and-effect relationships. The experience gained shows that much of what could not be known before is gradually becoming known with the development of quantitative methods and the improvement of technical means. Therefore, in the analysis, random effects, as well as the effects of unknown causes, are not rejected.

Topic 2. One-factor linear econometric model

2.1. Paired linear regression model

The functional and technical capabilities of modern computer technology have turned multidimensional statistical analysis from a theoretical section of mathematical statistics into a powerful tool for applied research of socio-economic phenomena and processes characterized by the multidimensionality of the parameters that describe them. The basis of multivariate statistical analysis is the methods of regression analysis.

Regression is understood as a one-way stochastic dependence of one random variable on another or several other random variables. Thus, regression establishes a correspondence between random variables. For example, when determining the dependence of the amount of tax revenues (y) on the tax rate (x), we are talking about determining a one-way relationship, that is, about regression. Both variables are random. Each value of x corresponds to a set of values of y and vice versa, each value of y corresponds to a set of values of x. Thus, we are dealing with a statistical distribution of x and y values. Based on these distributions, we must find a stochastic relationship between y and x. One-sided stochastic dependence is expressed using a function, which, unlike a strict mathematical dependence, is called a regression function or simply regression.

Often there are relationships between two or more variables for which a logical interpretation is possible in only one direction, and as a result, it is advisable to find only one regression function.

The regression function formally establishes correspondence between variables, although they may not be in a causal relationship. In this case, so-called false regressions may occur, which have no practical value and no meaning at all. Therefore, when constructing regression equations, one should always proceed from real problems that have an applied nature.

Let's move on to the classification of regressions according to the number of variables in the model and forms of dependence. Simple (pairwise) and multiple (multifactorial) regressions are distinguished by the number of variables entered into the regression equation. Regarding the form of dependence, models are divided into linear and non-linear regression.

Simple (pairwise) linear regression establishes a linear relationship between two variables. At the same time, one of the variables (y) is considered a dependent variable (exogenous, regressor, result variable, response) and is considered as a function of the second (x) independent variable (endogenous, explanatory, regressor).

For the general case, a simple linear model will be written as follows:

$$y = \alpha + \beta x + u \qquad (2.1)$$

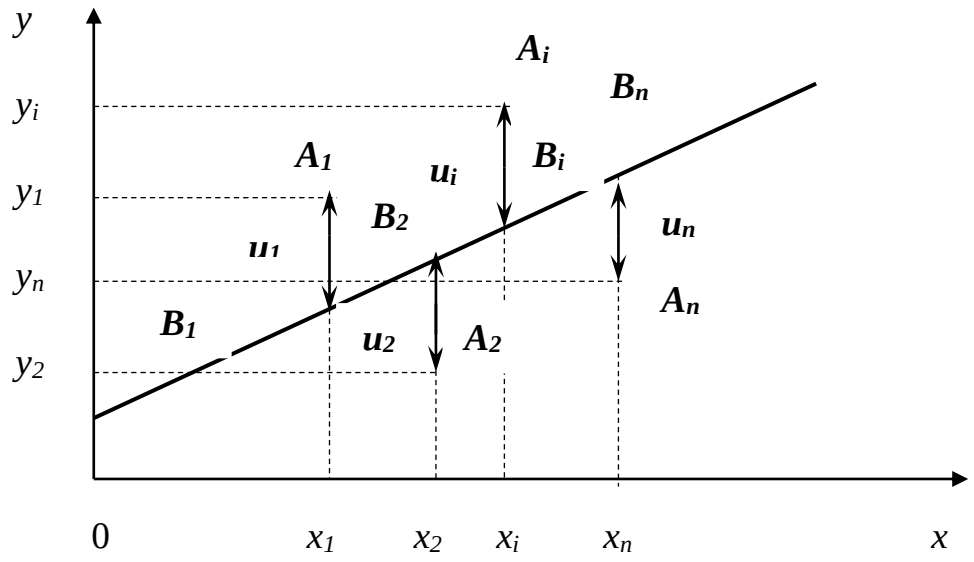
Value $y=\{y_1, y_2,.., y_n\}$ consists of two components:

1) non-random component $\alpha+\beta x$, де $x=\{x_1,x_2,.., x_n\}$, α і β - equation parameters;

2) random component (disturbances, errors) $u=\{u_1,u_2,..,u_n\}$.

Let's consider the geometric interpretation of the combination of these two components (Fig. 2.1).

Indicators $x_1, x_2,..., x_n$ – these are the hypothetical values of the explanatory variable. If the ratio between y and x were the same, then the corresponding values of y would be represented by points $B_1, B_2, ...,B_n$ on the same straight line. The presence of a random perturbation term leads to the fact that in reality the value of y is obtained by another. Let's mark the real values of y at the corresponding values of x on the graph with the help of points $A_1, A_2, ..., A_n$.



Pic. 2.1. The actual dependence between y and x

Actual parameter values α and β random component u and position of points $B_1, B_2, ..., B_n$ unknown. The task of regression analysis is to find estimates α and β and, as a result, in the position of a straight line by points $A_1,.A_2,..., A_n$.

If the random variable u (disturbance) would be absent, then points A_i would coincide with the points B_i and accurately showed the position of the line. In this case, it would be enough to simply construct this straight line and determine the values of α and β.

Let's consider the main reasons for the existence of the disturbance.

1. Exclusion of explanatory variables in the model. Establishing a relationship between only the two factors y and x is a great simplification. In fact, there are other factors that significantly affect the performance indicator, which are not taken into account or cannot be taken into account in formula (2.1). The influence of these factors leads to the fact that the true points lie outside the straight line. By combining all such possible components of influence on the performance indicator, we will get the value of u. For example, when studying the dependence of national income on capital investments, the disturbed variable u would include the influence on national income of the following factors: the number of workers employed in the field of production, labor productivity, the use of fixed assets and other random factors. If we knew exactly which variables should be included in the model, and had the opportunity to accurately measure them, then we could build equations based on them and thereby exclude the corresponding element of disturbance.

2. Aggregation of variables. In many cases, the proposed model is an attempt to combine a certain number of economic indicators. Since individual economic indicators have different characteristic parameters, the attempt to determine the relationship between them is of an approximation nature. The existing discrepancy is attributed to the existing disturbance.

3. Incorrect description of the structure of the model. The structure of the model may be described incorrectly or not completely correctly.

4. Incorrect functional specification. The functional relationship between y and x may not be mathematically defined correctly. For example, the actual relationship is not linear, but may be more complex. However, the use of even the most appropriate formula is somewhat approximate, and therefore the existing

discrepancy makes its corrections in the residual term.

5. Measurement errors. If there are errors in the measurements of one or more interrelated variables, then the values found will not correspond to the exact relationship, and the existing discrepancy will contribute to the structure of the perturbed variable.

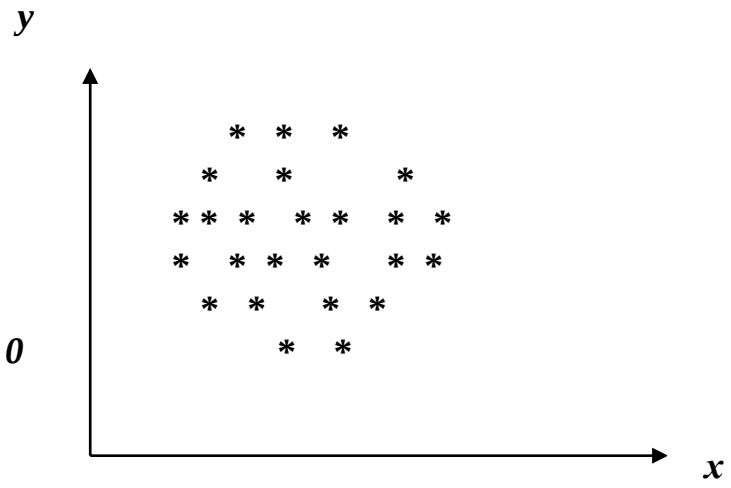
Therefore, the disturbance is a total manifestation of the factors listed above.

2.2. Scatter diagram of the regression function

To analyze the dependence between two variables, a scatter diagram is used, which is a graphical form of information presentation in a rectangular coordinate system. The value of the independent variable (x) is marked on the abscissa axis, and the value of the dependent variable (y) on the ordinate axis. The result of each observation (x_i, y_i) of some economic process is displayed by a point on the plane. The set of these points forms a cloud that shows the relationship between the two variables. A scatter diagram is a geometric form of systematization of the information base of the research process.

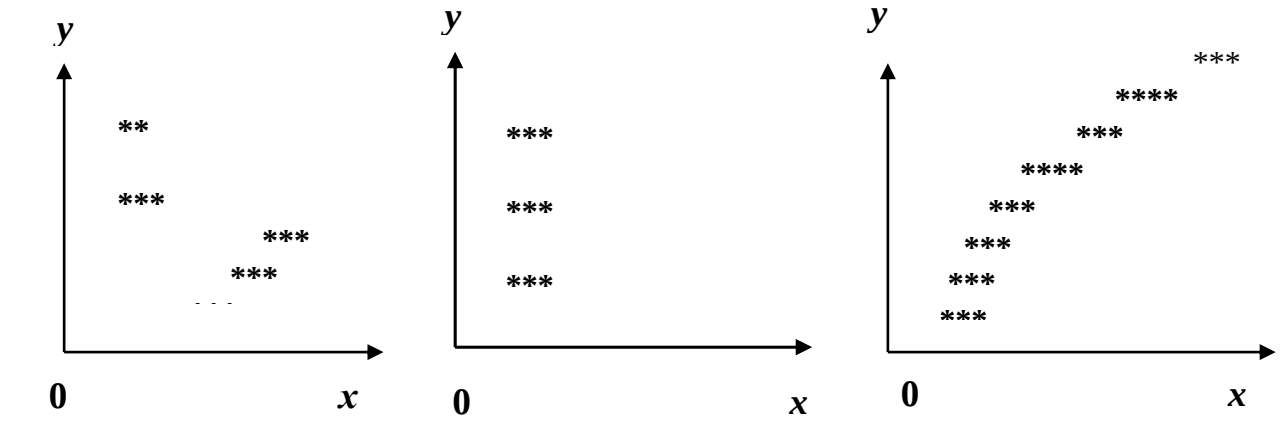
By the width of the spread of points, it is possible to draw a conclusion about the closeness of the connection of the population. If the points are placed close to each other (in the form of a narrow strip), it can be argued that there is a relatively close connection. If the points on the diagram are widely scattered, then there is a weak relationship between the variables (Fig. 2.2).

A partial solution to the main problem of regression analysis for the case of paired dependence can be achieved by constructing a scatter diagram. Fig. 2.3 shows the main forms of dependencies.

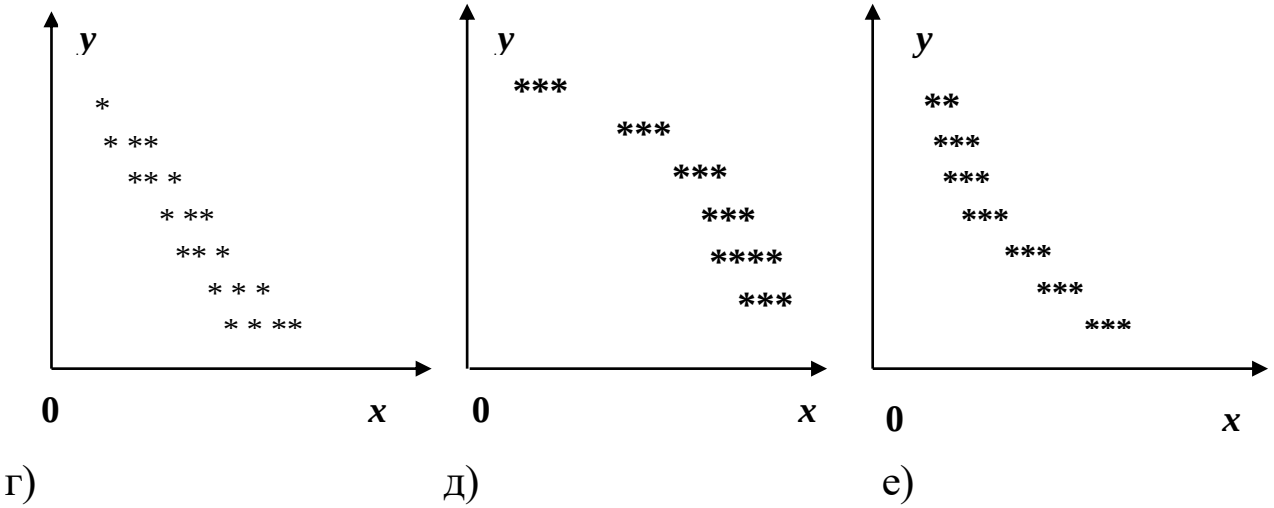


Pic. 2.2. Scatter plot in the case of no communication

Positive regression



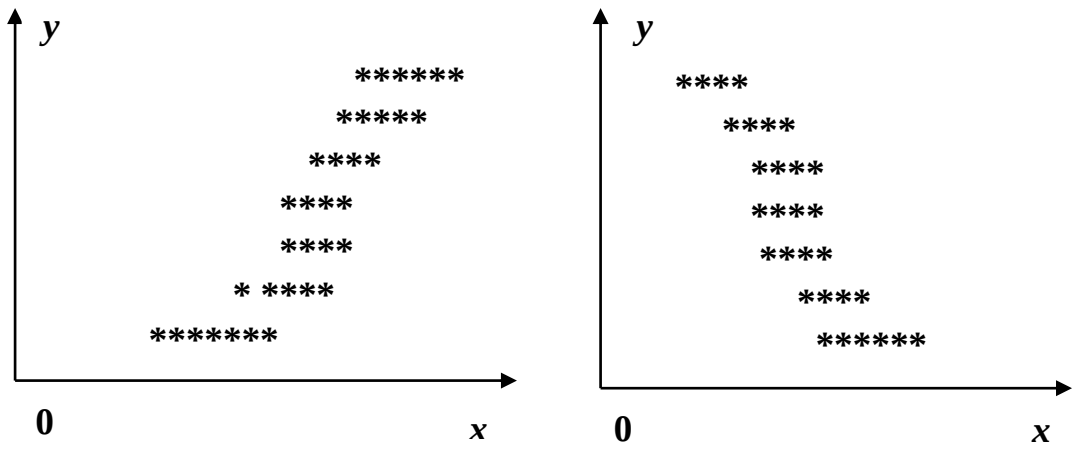
Negative regression



Pic. 2.3. Basic forms of regressions

Therefore, positive linear and non-linear and negative linear and non-linear regression are distinguished. For example, such a situation will occur when studying the dependence of the output volume on the cost of the main production assets. A positive equally accelerated increasing regression (Fig. 2.3 b) exists, for example, between income tax and wages. A positive equal-decelerated growing regression (Fig. 2.3 c) can take place when establishing the dependence of the level of labor productivity on the length of service.

A negative linear regression (Fig. 2.3. d) shows a uniform decline of the function, for example, the dependence of the number of enterprises in the region that will produce this type of product on the tax rate. In fig. 2.2d and 2.2e schematically present the corresponding situations of the relationship of the negative equally accelerated and equally decelerated downward regression. Very often, the given varieties of regressions are not found in their pure form, but in combination with each other, as can be seen in fig. 2.3. Regressions of this type are called combined forms.



Pic. 2.4. Combined forms of regressions

The given diagrams show that each value of the explanatory variable corresponds to the distribution of the values of the dependent variable and vice versa.

Connection is sought based on these distributions. It is important not only to indicate the general trend of changes in the dependent variable, but also to find out what the effect of the main factors-arguments on the dependent variable would be, if the others (secondary, secondary) did not change and were at the same average level. For this, the regression function is determined in the form of a mathematical equation of one form or another. The process of finding the regression function is called equalization of individual values of the dependent variable. Constructing a regression and determining the impact of explanatory variables on the dependent variable is the second task of regression analysis.

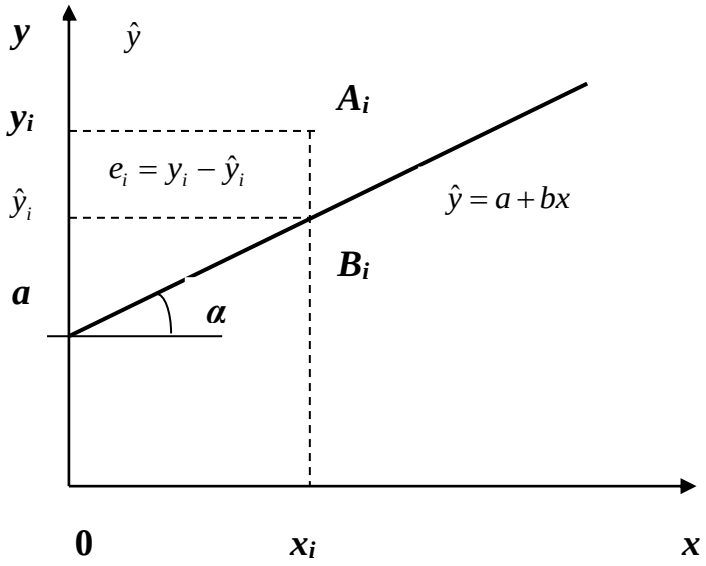
Based on the appearance of the cluster of points, it is possible to hypothesize the linearity or non-linearity of the relationship between the variables. So, on diagram 2.3 a, d, we have clearly expressed linear trends of clustering of points. Let's try to approximate the dependencies depicted in these diagrams with a linear regression function. Of course, these trends only exist on average. They are disturbed by the deviation of individual points The deviation from the straight line is explained by the influence of other unaccounted factors.

A scatter diagram allows visual analysis of empirical data. If the geometric representation of the dependence of three variables in space is still possible, albeit difficult, then with a larger number of variables this representation is impossible.

Let's assume that by the appearance of the scatter diagram, we have established the linear nature of the dependence of the averaged values of the resulting variable. Let's express this dependence with the help of the linear regression estimation function:

$$\hat{y} = a + bx, \tag{2.2}$$

where a та b respectively are parameter estimates, are parameter estimates α та β equation (2.1). Sign "^" over y means the estimate of the dependent variable obtained from equation (2.2) under some averaged conditions. So, simple regression means one-way stochastic dependence of the outcome variable on one explanatory variable.



Pic. 2.5. Regression line and its parameters

The constant value *a* determines the point of intersection of the direct regression with the ordinate axis (Fig. 2.4) and is the average value at the point $x_0 = 0$. It is clear that the economic interpretation of *a* is not only difficult, but also impossible. The value *a* in the regression equation only performs the function of equalization and has the dimension y. At the same time, it should be noted that thanks to the constant *a*, the regression function is infallible. The regression equation is interpreted only in the region of the cluster of points, and as a result, only between the smallest and the largest value of the variable *x* that is observed. Of more practical interest is the economic meaning of the values *b* and $\hat{y}$.

Coefficient *b* characterizes the slope of the straight line to the abscissa axis. We denote by α the angle that the straight line makes with the abscissa axis. Then $b=tg\ \alpha$. The regression coefficient *b* is a measure of the dependence of the variable *y* on *x* or a measure of the effect revealed by the change of *x* on *y* . According to

equation (2.2), b determines the average value of the change in the performance indicator when the explanatory variable x changes by one unit. The sign b determines the direction of this change, and the dimension of this coefficient is the ratio of the dimension of the dependent variable to the dimension of the explanatory variable.

After determining the numerical estimates of the parameters, it is possible according to the equation (2.2) calculate the value $\hat{y}_i$ for each value of the explanatory variable x_i. This value is called estimated.

With a linear function, the set of estimated values forms a regression line. As noted earlier, due to the random influence of extraneous factors for each value x_i several empirical values may be observed y_i, that is, to each value x corresponds to the distribution of values of the variable y. The value of the regression function $\hat{y}_i$ are thus estimates of the mean values of the variable y for each fixed value of the variable x.

An economic interpretation follows from this $\hat{y}_i$. Regression values $\hat{y}_i$ show the mean value of the dependent variable y at the given x_i explanatory variable in the assumption that the only reason for the change y is a variable x, and is a random perturbed variable u assumed a value equal to zero. The scatter of the observed values of the variable y around $\hat{y}_i$ caused by the influence of a number of unaccounted factors. The difference between the empirical value y_i and calculating $\hat{y}_i$ let's call the residual, which gives a numerical estimate of the value of the perturbed variable u (pic. 2.5). So, the numerical value e is defined as $y_i - \hat{y}_i = e_i, i = \overline{1,n}$. It is clear that the smaller the value e_i the more successfully chosen straight line.

2.3. The method of least squares

Task. For ten enterprises of the region, for a certain conditional period, the numerical values of two economic indicators are known: gross production ***y*** (million dollars) and the cost of the main production assets ***x*** (million dollars), (Table 1). To study the characteristics of the influence of the cost of the main production assets (***x***) on the output of gross products (***y***) of the enterprise:

1. Make a model specification.
2. Find statistical estimates of the parameters of the linear regression equation and construct an estimation line.
3. Calculate the value of the point estimate of the dispersion of disturbances.
4. For the level of significance $\alpha = 0{,}05$ check the significance of the regression coefficients α_0 and α_1.
5. Find the confidence intervals of the regression coefficients with reliability $\gamma = 0{,}95$.
6. Calculate the values of the sample: coefficient of determination, coefficient of correlation.
7. Find and construct the confidence interval of the regression function with reliability $\gamma = 0{,}95$.
8. Check the adequacy of the constructed econometric model. If the model is adequate, then find the forecast value of the gross

output for the value of the main production assets in the amount of UAH 12 million, with reliably $\gamma = 0{,}95$ build a confidence interval for this forecast value.

Table 1

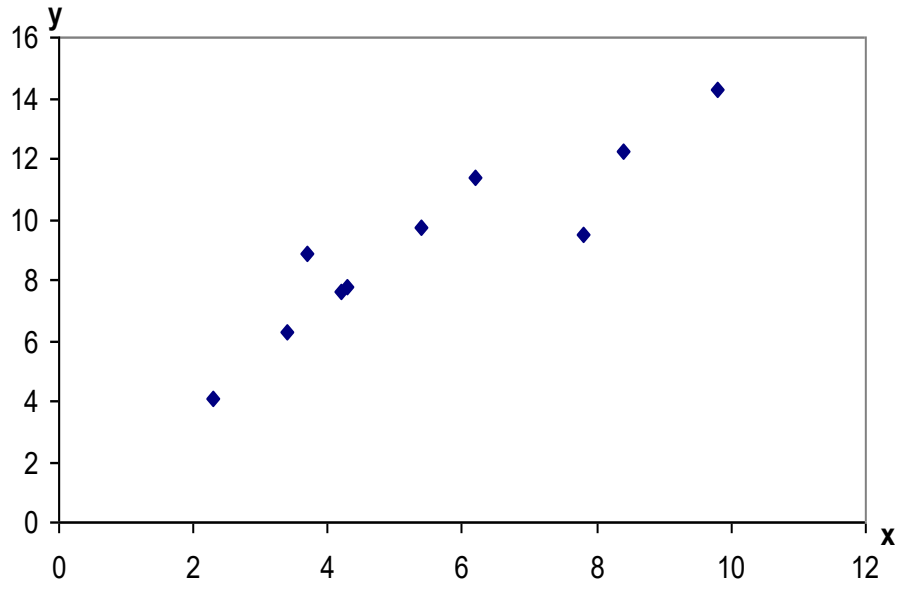
№ firms	Gross output, mln. dol., y_i	Main production assets, mln. dol., x_i
1	4,1	2,3
2	6,3	3,4
3	7,6	4,2
4	8,9	3,7
5	7,8	4,3
6	9,7	5,4
7	11,4	6,2
8	9,5	7,8
9	12,2	8,4
10	14,3	9,8

♦ ***Solution.***

1. In order to determine the form of dependence between two variables, a scatter diagram is used, which is a graphical representation of information in a rectangular coordinate system. The abscissa indicates the value of the independent variable (x), on the y-axis – the value of the dependent variable (y). The result of each observation (x_i, y_i) of some economic process is represented by a point on the plane. The set of these points forms a cloud that shows the relationship between the two variables.

By the width of the spread of points, it is possible to draw a conclusion about the closeness of the connection of the population. If the points are placed close to each other (in the form of a narrow strip), then it can be argued that there is a relatively close connection. If the points on the plot are widely scattered, there is a weak or no relationship between the variables.

Based on the appearance of the cluster of points, it is possible to hypothesize the linearity or non-linearity of the relationship between the variables. Thus, on diagram 1 (a, г), we have clearly expressed linear trends of clustering of points. Let's build a scatter diagram of the dependence of gross output (y) on the cost of the enterprise's main production assets (x):



The placement of points on the scatter diagram makes it possible to make an assumption about the existence of a linear form of the relationship in the form of a function:

$$\hat{y} = a_0 + a_1 x, \qquad (1)$$

where $\hat{y}$ – estimated volume of gross output, mln. dol.; x – the cost of

the main production assets, mln. dol.

2. Statistical estimates of the parameters of the linear regression equation can be found by the method of least squares (LSM), which is based on the requirement to minimize the sum of squares of the deviations of the empirical values of the variable y from the values calculated by the equation of the straight line. With its help, such estimates of the parameters of the regression equation are found that minimize the selected measure of dispersion. These estimates can be found by the system of normal

$$\begin{cases} na_0 + \left(\sum_{i=1}^{n} x_i\right) a_1 = \sum_{i=1}^{n} y_i, \\ \left(\sum_{i=1}^{n} x_i\right) a_0 + \left(\sum_{i=1}^{n} x_i^2\right) a_1 = \sum_{i=1}^{n} x_i y_i. \end{cases} \quad (2)$$

Dividing each of the equations by n, we will get:

$$\begin{cases} a_0 + \bar{x} a_1 = \bar{y}, \\ \bar{x} a_0 + \overline{x^2} a_1 = \overline{xy}, \end{cases} \quad (3)$$

where $\bar{x} = \sum x_i / n$, $\bar{y} = \sum y_i / n$, $\overline{xy} = \sum x_i y_i / n$, $\overline{x^2} = \sum x_i^2 / n$.

Solution of the system (3) find by Kramer's rule:

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \left(\bar{x}\right)^2}, \quad a_0 = \bar{y} - a_1 \bar{x}. \quad (4)$$

The value of the same estimates can be found using formulas due to the deviation from the average:

$$a_1 = \frac{\sum_{i=1}^{n}(\Delta x_i \Delta y_i)}{\sum_{i=1}^{n}(\Delta x_i)^2}. \qquad (5)$$

The value of the assessment *a* we find from the formula (1.7):

$$a_0 = \bar{y} - a_1 \bar{x}, \qquad (6)$$

where $\Delta y_i = y_i - \bar{y}$ and $\Delta x_i = x_i - \bar{x}$ – this is the variance of the variables *y* and *x* from their average values.

To simplify calculations when finding estimates a_0 and a_1 let's build a table of parameters of the econometric model:

Table 2

№	y_i	x_i	x_i^2	y_i^2	$xi \cdot yi$	Δx_i	$(\Delta x_i)^2$	Δy_i	$\Delta x_i \cdot \Delta y_i$
1	4,1	2,3	5,29	16,81	9,43	-3,25	10,56	-5,08	16,51
2	6,3	3,4	11,56	39,69	21,42	-2,15	4,62	-2,88	6,19
3	7,6	4,2	17,64	57,76	31,92	-1,35	1,82	-1,58	2,13
4	8,9	3,7	13,69	79,21	32,93	-1,85	3,42	-0,28	0,52
5	7,8	4,3	18,49	60,84	33,54	-1,25	1,56	-1,38	1,73
6	9,7	5,4	29,16	94,09	52,38	-0,15	0,02	0,52	-0,08
7	11,4	6,2	38,44	129,96	70,68	0,65	0,42	2,22	1,44
8	9,5	7,8	60,84	90,25	74,1	2,25	5,06	0,32	0,72
9	12,2	8,4	70,56	148,84	102,48	2,85	8,12	3,02	8,61
10	14,3	9,8	96,04	204,49	140,14	4,25	18,06	5,12	21,76
Сума	**91,8**	**55,5**	**361,71**	**921,94**	**569,02**	**0**	**53,69**	**0**	**59,53**

Let's calculate the average values:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{55,5}{10} = 5,55; \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{91,8}{10} = 9,18.$$

$$\overline{x^2} = \frac{\sum_{i=1}^{n} x_i^2}{n} = \frac{361,71}{10} = 36,17; \qquad \overline{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} = \frac{569,02}{10} = 56,9.$$

Let's write down the system (3):

$$\begin{cases} a_0 + 5,55a_1 = 9,18, \\ 5,55a_0 + 36,17a_1 = 56,9, \end{cases}$$

and solve it using the formulas (4):

$$a_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{56,9 - 5,55 \cdot 9,18}{36,17 - 5,55^2} = \frac{5,95}{5,37} = 1,11,$$

$$a_0 = \bar{y} - a_1 \bar{x} = 9,18 - 1,11 \cdot 5,55 = 3,02.$$

Let's find the same estimates using the formulas due to the deviation from the average. First, let's calculate the deviation of the variables from their average values:

$$\Delta x_1 = x_1 - \bar{x} = 2,3 - 5,55 = -3,25;$$
$$\Delta x_2 = x_2 - \bar{x} = 3,4 - 5,55 = -2,15;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\Delta x_{10} = x_{10} - \bar{x} = 9,8 - 5,55 = 4,25.$$

$$\Delta y_1 = y_1 - \bar{y} = 4,1 - 9,18 = -5,08;$$
$$\Delta y_2 = y_2 - \bar{y} = 6,3 - 9,18 = -2,88;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\Delta y_{10} = y_{10} - \bar{y} = 14,3 - 9,18 = 5,12.$$

Then $$a_1 = \frac{\sum_{i=1}^{n}(\Delta x_i \Delta y_i)}{\sum_{i=1}^{n}(\Delta x_i)^2} = \frac{59{,}53}{53{,}69} = 1{,}11,$$

$$a_0 = \bar{y} - a_1 \bar{x} = 9{,}18 - 1{,}11 \cdot 5{,}55 = 3{,}02, \quad .$$

So, we obtained the estimation equation of the econometric model

$$\hat{y} = 3{,}02 + 1{,}11x.$$

Let's construct this estimation straight line.

3. Unbiased point estimate S_u^2 of the unknown variance of disturbances σ_u^2 we find by the formula:

$$S_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2.$$

First, let's calculate the estimated values $\hat{y}_i$ according to the estimation equation of the econometric model $\hat{y} = 3{,}02 + 1{,}11x$, then deviation $u_i = y_i - \hat{y}_i$ and $u_i^2 = \left(y_i - \hat{y}_i\right)^2$.

Table 3

№	y_i	x_i	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	4,1	2,3	5,58	-1,48	2,18
2	6,3	3,4	6,80	-0,50	0,25
3	7,6	4,2	7,68	-0,08	0,01
4	8,9	3,7	7,13	1,77	3,14
5	7,8	4,3	7,79	0,01	0,00
6	9,7	5,4	9,01	0,69	0,47
7	11,4	6,2	9,90	1,50	2,25
8	9,5	7,8	11,67	-2,17	4,73
9	12,2	8,4	12,34	-0,14	0,02
10	14,3	9,8	13,89	0,41	0,17
Сума	**91,8**	**55,5**	**91,8**	**0**	**13,2**

Using the obtained calculations, we get:

$$S_u^2 = \frac{1}{10-2} \cdot 13{,}2 = 1{,}65.$$

4. Significance of regression coefficients a_0 та a_1 we check using the null hypothesis $H_0(\alpha_m = 0)$, the content of which and alternative to it $H_1(\alpha_m \neq 0)$ is as follows: if the inequality holds

$$\left|\frac{a_m}{S_{a_m}}\right| > t_{кр.},$$

where $t_{кр.} = t_{двост.кр.}(\alpha, n-2)$, $\alpha = 1-\gamma$, – critical point of the Student's distribution, then the hypothesis is accepted at the level of significance α the hypothesis H_1 is accepted, that is, it is considered that $\alpha_m \neq 0$.

Values S_{a_0} and S_{a_1} we will find using the formulas:

$$S_{a_0} = \sqrt{\frac{S_u^2 \overline{x^2}}{n\sigma_x^2}} = \frac{S_u}{\sigma_x}\sqrt{\frac{\overline{x^2}}{n}} = \frac{\sqrt{1{,}65}}{\sqrt{\overline{x^2} - (\overline{x})^2}}\sqrt{\frac{\overline{x^2}}{10}} = \frac{\sqrt{1{,}65}}{\sqrt{36{,}17 - (5{,}55)^2}}\sqrt{\frac{36{,}17}{10}} = 1{,}05$$

;

$$S_{a_1} = \sqrt{\frac{S_u^2}{n\sigma_x^2}} = \frac{S_u}{\sigma_x\sqrt{n}} = \frac{\sqrt{1{,}65}}{\sqrt{10[36{,}17 - (5{,}55)^2]}} = 0{,}17.$$

Then the empirical values of the criterion:

$$\left|\frac{a_0}{S_{a_0}}\right| = \frac{3{,}02}{1{,}05} = 2{,}88, \qquad \left|\frac{a_1}{S_{a_1}}\right| = \frac{1{,}11}{0{,}17} = 6{,}53.$$

Critical point for the two-sided critical region

$t_{кр.} = t_{двост.}(\alpha, k)$ at the values $\alpha = 0{,}05$, $k = n - 2 = 8$ we find from the table of critical values of the Student's distribution $t_{кр.} = 2{,}306$.

Since $2{,}88 > t_{кр.} = 2{,}306$ and $6{,}53 > t_{кр.} = 2{,}306$, then at the level of significance $\alpha = 0{,}05$ we conclude that $\alpha_0 \neq 0$ and $\alpha_1 \neq 0$.

5. Confidence intervals with reliability γ for unknown regression parameters a_0 and a_1 have the appearance:

$$a_m - t_m(\gamma, k)S_{a_m} < \alpha_m < a_m + t_m(\gamma, k)S_{a_m},$$

where $m = 0{,}1$, $t_m = t_m(\gamma, k)$ – root of equation $P(|t_m| < t) = \gamma$, t_0 and t_1 – random variables distributed according to Student's law.

In our case $\gamma = 0{,}95$, the number of degrees of freedom $k = n - 2 = 8$.

We find by the table $t_0(0{,}95;8) = t_1(0{,}95;8) = 2{,}306$. Then, taking into account the found values $S_{a_0} = 0{,}1055$, $S_{a_1} = 0{,}1132$ we will get:

$$3{,}02 - 2{,}306 \cdot 1{,}05 < \alpha_0 < 3{,}02 + 2{,}306 \cdot 1{,}05,$$

$$1{,}11 - 2{,}306 \cdot 0{,}17 < \alpha_1 < 1{,}11 + 2{,}306 \cdot 0{,}17$$

or

$$0{,}6 < \alpha_0 < 5{,}44,$$

$$0{,}72 < \alpha_1 < 1{,}5.$$

6. Coefficient of determination R^2 we find by the formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n\sigma_y^2}.$$

From the table 3 $\sum_{i=1}^{10} (y_i - \hat{y}_i)^2 = 13{,}02.$

Calculate σ_y^2, using the calculations of item 2:

$$\sigma_y^2 = \overline{y^2} - (\overline{y})^2 = 921{,}94/10 - (9{,}18)^2 = 7{,}92.$$

Then

$$R^2 = 1 - \frac{13{,}02}{10 \cdot 7{,}92} = 0{,}836.$$

Thus, the variation of the dependent variable Y on 83,6% explained by the variation of the explanatory variable.

The sample correlation coefficient can be found using the formula:

$$r = \sqrt{R^2} = \sqrt{0{,}836} = 0{,}914.$$

At the same time, the positive sign of this number is chosen due to the fact that $a_1 > 0$.

7. In order to construct a confidence interval for the regression function, it is necessary to place points with coordinates on the coordinate plane $\{x_i; \hat{y}_i - t(\gamma, n-2) S_{\hat{y}_i}\}$, $i = \overline{1, n}$ and connect adjacent (by index) points with straight lines, and then carry out a similar procedure for points

$\{x_i; \hat{y}_i + t(\gamma, n-2) S_{\hat{y}_i}\}.$

Let's calculate the value $S_{\hat{y}_i}$ according to the formula:

$$S_{\hat{y}_i} = S_u \sqrt{\left[1 + \frac{\left(x_i - \bar{x}\right)^2}{\sigma_x^2}\right]\frac{1}{n}}.$$

We will use the value found in point 3

$S_u = \sqrt{1{,}65} = 1{,}28$ та $\sigma_x^2 = \overline{x^2} - \left(\bar{x}\right)^2 = 36{,}17 - 5{,}55^2 = 5{,}37$,

we will get:

$$S_{\hat{y}_1} = 1{,}28\sqrt{\left[1 + \frac{\left(2{,}3 - 5{,}55\right)^2}{5{,}37}\right]\frac{1}{10}} = 0{,}698;$$

$$S_{\hat{y}_1} = 1{,}28\sqrt{\left[1 + \frac{\left(3{,}4 - 5{,}55\right)^2}{5{,}37}\right]\frac{1}{10}} = 0{,}552;$$

$S_{\hat{y}_3} = 0{,}469;$ $S_{\hat{y}_4} = 0{,}518;$ $S_{\hat{y}_5} = 0{,}46;$ $S_{\hat{y}_6} = 0{,}406;$

$S_{\hat{y}_7} = 0{,}421;$ $S_{\hat{y}_8} = 0{,}564;$ $S_{\hat{y}_9} = 0{,}462;$ $S_{\hat{y}_{10}} = 0{,}846.$

We find by the table $t(0{,}95;8) = 2{,}306$. Using the value $\hat{y}_i$ from table 3 and values $S_{\hat{y}_i}$, we will get the ordinates of the points of the lower limit of the confidence zone:

$\hat{y}_1 - tS_{\hat{y}_1} = 5{,}58 - 2{,}306 \cdot 0{,}698 = 3{,}97;$

$\hat{y}_2 - tS_{\hat{y}_2} = 6{,}8 - 2{,}306 \cdot 0{,}552 = 5{,}52;$

$\hat{y}_3 - tS_{\hat{y}_3} = 7{,}68 - 2{,}306 \cdot 0{,}469 = 6{,}6;$

$\hat{y}_4 - tS_{\hat{y}_4} = 7{,}13 - 2{,}306 \cdot 0{,}518 = 5{,}93;$

$\hat{y}_5 - tS_{\hat{y}_5} = 7{,}79 - 2{,}306 \cdot 0{,}46 = 6{,}73$

$\hat{y}_6 - tS_{\hat{y}_6} = 9{,}01 - 2{,}306 \cdot 0{,}406 = 8{,}08;$

$$\hat{y}_7 - tS_{\hat{y}_7} = 9{,}9 - 2{,}306 \cdot 0{,}421 = 8{,}93;$$

$$\hat{y}_8 - tS_{\hat{y}_8} = 11{,}67 - 2{,}306 \cdot 0{,}564 = 10{,}97;$$

$$\hat{y}_9 - tS_{\hat{y}_9} = 12{,}34 - 2{,}306 \cdot 0{,}462 = 10{,}86;$$

$$\hat{y}_{10} - tS_{\hat{y}_{10}} = 13{,}89 - 2{,}306 \cdot 0{,}846 = 11{,}94.$$

Then the ordinates of the points of the upper limit of the confidence zone take the following:

$$\hat{y}_1 + tS_{\hat{y}_1} = 5{,}58 + 2{,}306 \cdot 0{,}698 = 7{,}18;$$

$$\hat{y}_2 + tS_{\hat{y}_2} = 6{,}8 + 2{,}306 \cdot 0{,}552 = 8{,}07;$$

$$\hat{y}_3 + tS_{\hat{y}_3} = 7{,}68 + 2{,}306 \cdot 0{,}469 = 8{,}76;$$

$$\hat{y}_4 + tS_{\hat{y}_4} = 7{,}13 + 2{,}306 \cdot 0{,}518 = 8{,}32;$$

$$\hat{y}_5 + tS_{\hat{y}_5} = 7{,}79 + 2{,}306 \cdot 0{,}46 = 8{,}86$$

$$\hat{y}_6 + tS_{\hat{y}_6} = 9{,}01 + 2{,}306 \cdot 0{,}406 = 9{,}95;$$

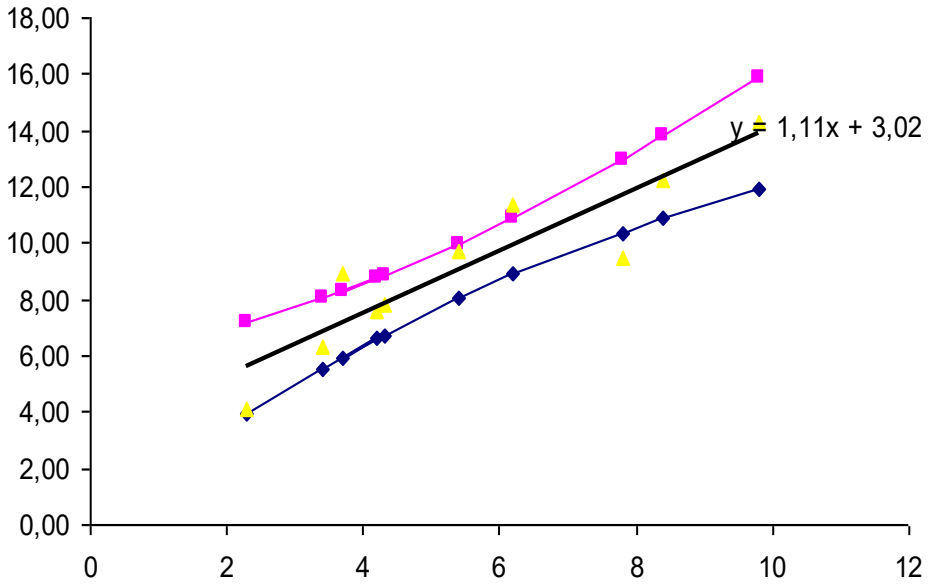
$$\hat{y}_7 + tS_{\hat{y}_7} = 9{,}9 + 2{,}306 \cdot 0{,}421 = 10{,}87;$$

$$\hat{y}_8 + tS_{\hat{y}_8} = 11{,}67 + 2{,}306 \cdot 0{,}564 = 12{,}98;$$

$$\hat{y}_9 + tS_{\hat{y}_9} = 12{,}34 + 2{,}306 \cdot 0{,}462 = 13{,}98;$$

$$\hat{y}_{10} + tS_{\hat{y}_{10}} = 13{,}89 + 2{,}306 \cdot 0{,}846 = 15{,}84.$$

Confidence interval with reliability 0,95 for the regression function has the form:



8. The adequacy of the constructed econometric model can be checked using the coefficient of determination. If its value is close to one, it can be considered that the obtained econometric model is adequate. In this case, the change in the value of the outcome variable *y* depends linearly on the change in the explanatory variable *x*, and not due to the influence of random factors. If the value of the coefficient of determination is close to zero, then the model is considered inadequate, that is, there is no linear relationship between *y* and *x*. If the value of the coefficient of determination is unclear, that is, close to 0.5, then the Fisher *F*. test is used to check the adequacy of the econometric model.

We calculate the empirical value of the *F*. parameter using the formula:

$$F_{емп} = \frac{R^2(n-m-1)}{m(1-R^2)},$$

where m – the number of independent variables (for simple regression m=1).

After that, we find the value from the table $F_{кр}$ – critical value *F*-Fisher distribution with з (k_1=m=1, k_2=n-m-1) degrees of freedom and level of significance α. For example, α=0,05, then we can be wrong in 5 % cases, and in 95 % of cases our conclusions will be correct.

If the value calculated by us $F_{емп}$>$F_{кр}$, then the econometric model built by us is adequate to reality.

In the opposite case, that is, if calculated $F_{емп} \leq F_{кр}$, then in this case the econometric model we built is inadequate to the real reality.

Let's calculate:

$$F_{емп} = \frac{R^2(n-m-1)}{m(1-R^2)} = \frac{0{,}836 \cdot (10-1-1)}{1 \cdot (1-0{,}836)} = 40{,}78.$$

Let's find the tabular value of this criterion ($F_{кр.}$) for the reliability level p=0,95 and the number of degrees of freedom k_1=m=1, k_2=n-m-1=10-1-1=8:

$$F_{кр} = 5{,}32.$$

Since $F_{емп.} > F_{кр.}$, then the estimation equation of the econometric model obtained by us

$$\hat{y} = 3{,}02 + 1{,}11x$$

adequate to the real reality and on its basis it is possible to make forecasts, i.e. predict the ways of development of the studied phenomena and processes for the near future. The forecast can be point or interval.

Point forecast for the next one n+1 we get the period when we substitute the value of the explanatory variable x_{n+1} into the estimation equation of the econometric model. The forecast value of the gross production output for the cost of the main production assets in the amount of 12 million dollars will be:

$$\hat{y}_{n+1} = 3{,}02 + 1{,}1 \cdot 12 = 16{,}22.$$

An interval forecast is an interval in which with a given probability $p=1-\alpha$ will get a valid value of the result variable y.

Confidence interval for the predicted value y_{n+1} with reliability $\gamma = 0{,}95$ looks like:

$$\hat{y}_{n+1} - t(\gamma; n-2) S_{u_{n+1}} < y_{n+1} < \hat{y}_{n+1} + t(\gamma; n-2) S_{u_{n+1}},$$

where $$S^2_{u_{n+1}} = S^2_u \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{n\sigma_x^2} \right].$$

The lower limit of this interval is called a pessimistic forecast, and the upper limit is called an optimistic forecast. First for x_{n+1}=12 find:

$$S_{u_{n+1}} = S_u \sqrt{1+\frac{1}{n}+\frac{\left(x_{n+1}-\bar{x}\right)^2}{n\sigma_x^2}} = 1{,}28\sqrt{1+\frac{1}{10}+\frac{(12-5{,}55)^2}{10\cdot 5{,}37}} = 1{,}75.$$

Then the desired confidence interval has the form

$$16{,}22 - 2{,}306\cdot 1{,}75 < y_{n+1} < 16{,}22 + 2{,}306\cdot 1{,}75$$

or

$$12{,}18 < y_{n+1} < 20{,}26.$$

Topic. Classical linear multivariate model

Construction of an econometric model with three variables by the OLS method

Task. Let consider the sample of data Y, X_1, X_2 *in the next table*

Table

Y_i	*4*	*5*	*6*	*8*	*11*	*11*	*12*	*12*	*13*	*14*
X_1	*3*	*4*	*5*	*7*	*9*	*11*	*10*	*12*	*11*	*12*
X_2	*7*	*9*	*11*	*12*	*12*	*15*	*18*	*21*	*22*	*24*

Determine the grades $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ *using OLS for linear econometric model.*

Solution. *Find* $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} \\ \sum_{i=1}^{n} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} X_{1i} \\ \sum_{i=1}^{n} Y_i X_{2i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2 \end{cases}$$

№	Y_i	X_{1i}	X_{2i}	X_{1i}^2	X_{2i}^2	$Y_i X_{1i}$	$Y_i X_{2i}$	$X_{1i} X_{2i}$
1	4	3	7	9	49	12	28	21

2	5	4	9	16	81	20	45	36
3	6	5	11	25	121	30	66	55
4	8	7	12	49	144	56	96	84
5	11	9	12	81	144	99	132	108
6	11	11	15	121	225	121	165	165
7	12	10	18	100	324	120	216	180
8	12	12	21	144	441	144	252	252
9	13	11	22	121	484	143	286	242
10	14	12	24	144	576	168	336	288
Σ	**96**	**84**	**151**	**810**	**2589**	**913**	**1622**	**1431**

$$\begin{cases} 96 = 10\hat{\beta}_0 + 84\hat{\beta}_1 + 151\hat{\beta}_2 \\ 913 = 84\hat{\beta}_0 + 810\hat{\beta}_1 + 1431\hat{\beta}_2 \\ 1622 = 151\hat{\beta}_0 + 1431\hat{\beta}_1 + 2589\hat{\beta}_2 \end{cases}$$

$$\begin{pmatrix} 10 & 84 & 151 \\ 84 & 810 & 1431 \\ 151 & 1431 & 2589 \end{pmatrix}$$

$$\begin{pmatrix} 0.848978 & -0.02401 & -0.03625 \\ -0.02401 & 0.053163 & -0.02798 \\ -0.03625 & -0.02798 & 0.017968 \end{pmatrix}$$

$$\begin{pmatrix} 0.848978 & -0.02401 & -0.03625 \\ -0.02401 & 0.053163 & -0.02798 \\ -0.03625 & -0.02798 & 0.017968 \end{pmatrix} \times \begin{pmatrix} 96 \\ 913 \\ 1622 \end{pmatrix} = \begin{pmatrix} 0.79 \\ 0.84 \\ 0.11 \end{pmatrix}$$

$\hat{\beta}_0 = 0.79$

$\hat{\beta}_1 = 0.84$

$\hat{\beta}_2 = 0.11$

$\hat{Y} = 0.79 + 0.84X_1 + 0.11X_2$ **.**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1)\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2 - \sum_{i=1}^{n}(Y_i - \bar{Y})(X_{2i} - \bar{X}_2)\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)}{\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2 - \left(\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{2i} - \bar{X}_2)\right)^2}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_{2i} - \bar{X}_2)\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2 - \sum_{i=1}^{n}(Y_i - \bar{Y})(X_{1i} - \bar{X}_1)\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(X_{1i} - \bar{X}_1)}{\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)^2\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)^2 - \left(\sum_{i=1}^{n}(X_{2i} - \bar{X}_2)(X_{1i} - \bar{X}_1)\right)^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}_1 - \hat{\beta}_2\bar{X}_2$$

№	Y i	X 1i	X 2i	$\bar{Y}$	$\bar{X}_1$	$\bar{X}_2$	$Y_i - \bar{Y}$	$X_{1i} - \bar{X}_1$	$X_{2i} - \bar{X}_2$	$(X_{1i} - \bar{X}_1)^2$	$(X_{2i} - \bar{X}_2)^2$	$(X_{1i} - \bar{X}_1) \times (X_{2i} - \bar{X}_2)$	$(Y_i - \bar{Y}) \times (X_{1i} - \bar{X}_1)$	$(Y_i - \bar{Y}) \times (X_{2i} - \bar{X}_2)$
1	4	3	7	9,6	8,4	15,1	-5,6	-5,4	-8,1	29,16	65,61	43,74	30,24	45,36
2	5	4	9				-4,6	-4,4	-6,1	19,36	37,21	26,84	20,24	28,06
3	6	5	11				-3,6	-3,4	-4,1	11,56	16,81	13,94	12,24	14,76
4	8	7	12				-1,6	-1,4	-3,1	1,96	9,61	4,34	2,24	4,96
5	11	9	12				1,4	0,6	-3,1	0,36	9,61	-1,86	0,84	-4,34
6	11	11	15				1,4	2,6	-0,1	6,76	0,01	-0,26	3,64	-0,14
7	12	10	18				2,4	1,6	2,9	2,56	8,41	4,64	3,84	6,96
8	12	12	21				2,4	3,6	5,9	12,96	34,81	21,24	8,64	14,16
9	13	11	22				3,4	2,6	6,9	6,76	47,61	17,94	8,84	23,46
10	14	12	24				4,4	3,6	8,9	12,96	79,21	32,04	15,84	39,16
Σ	**96**	**84**	**151**				**0**	**0**	**0**	**104,4**	**308,9**	**162,6**	**106,6**	**172,4**

$$\sum_{i=1}^{10}\left(Y_i-\bar{Y}\right)^2=(-5.6)^2+(-4.6)^2+(-3.6)^2+(-1.6)^2+(1.4)^2+(1.4)^2+(2.4)^2+(2.4)^2+(3.4)^2+(4.4)^2=$$
$$=31.36+21.16+12.96+2.56+1.96+1.96+5.76+5.76+11.56+19.36=114.4$$

$$\hat{\beta}_1=\frac{106.6\cdot 308.9-172.4\cdot 162.6}{104.4\cdot 308.9-(162.6)^2}\approx 0.8427$$
$$\hat{\beta}_2=\frac{172.4\cdot 104.4-106.6\cdot 162.6}{104.4\cdot 308.9-(162.6)^2}\approx 0.1145$$
$$\hat{\beta}_0=9.6-0.84\cdot 8.4-0.1145\cdot 15.1=0.7919$$

$$\hat{Y}=0.79+0.84X_1+0.11X_2.$$

Topic. *Econometric model with three variables: hierarchy of correlation coefficients*

Task. *Calculate the coefficients of hair, patial and multiplying correlation.*

Solution. $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$,

$Y = \alpha' + \beta' X_1$

$Y = \alpha'' + \beta'' X_2$

$X_1 = \alpha''' + \beta''' X_2$

$$: \; r_{YX_1} = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_{1i} - \bar{X}_1\right)}{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2 \sum \left(X_{1i} - \bar{X}_1\right)^2}}, \; r_{YX_2} = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)\left(X_{2i} - \bar{X}_2\right)}{\sqrt{\sum \left(Y_i - \bar{Y}\right)^2 \sum \left(X_{2i} - \bar{X}_2\right)^2}},$$

$$r_{X_1X_2} = \frac{\sum_{i=1}^{n} \left(X_{1i} - \bar{X}_1\right)\left(X_{2i} - \bar{X}_2\right)}{\sqrt{\sum \left(X_{2i} - \bar{X}_2\right)^2 \sum \left(X_{1i} - \bar{X}_1\right)^2}}$$

$$r_{YX_1} = \frac{106.4}{\sqrt{114.4 \cdot 104.4}} = \frac{106.4}{109.28} \approx 0.97, \quad r_{YX_2} = \frac{172.4}{\sqrt{114.4 \cdot 308.9}} = \frac{172.4}{187.98} \approx 0.92,$$

$$r_{X_1X_2} = \frac{162.6}{\sqrt{104.4 \cdot 308.9}} = \frac{162.6}{179.58} \approx 0.91$$

$$|r| = \begin{pmatrix} 1 & 0.97 & 0.92 \\ 0.97 & 1 & 0.91 \\ 0.92 & 0.91 & 1 \end{pmatrix}$$

$$r_{YX2.X1} = \frac{r_{YX2} - r_{YX1} \cdot r_{X1X2}}{\sqrt{1 - r_{YX1}^2} \cdot \sqrt{1 - r_{X1X2}^2}} = \frac{0.92 - 0.97 \cdot 0.91}{\sqrt{1 - (0.97)^2} \cdot \sqrt{1 - (0.91)^2}} = 0.4$$

$$r_{YX1.X2} = \frac{r_{YX1} - r_{YX2} \cdot r_{X1X2}}{\sqrt{1 - r_{YX2}^2} \cdot \sqrt{1 - r_{X1X2}^2}} = \frac{0.97 - 0.92 \cdot 0.91}{\sqrt{1 - (0.92)^2} \cdot \sqrt{1 - (0.91)^2}} = 0.83$$

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 .$$

$$R^2 = \frac{r_{YX1}^2 + r_{YX2}^2 - 2r_{YX1}r_{YX2}r_{X1X2}}{1 - r_{X1X2}^2}$$

$$R^2 = \frac{(0.97)^2 + (0.92)^2 - 2 \cdot 0.97 \cdot 0.92 \cdot 0.91}{1 - (0.91)^2} = 0.95$$

$$R = \sqrt{R^2} = \sqrt{0.95} = 0.97$$

Topic. A matrix approach to a linear multivariate model

Task

Let consider the statistical data y, x_1, x_2 in the table. Calculate econometric model in matrix form.

y	4	5	6	8	11	11	12	12	13	14
x_1	3	4	5	7	9	11	10	12	11	12
x_2	7	9	11	12	12	15	18	21	22	24

Find vector of estimation of coefficients $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix}$ and matrix of dispertion of estimation $\mathrm{var}\left(\hat{\beta}\right)$.

Solution.

$$\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_1 + \hat{\beta}_3 x_2$$

$$\hat{\beta} = \left[[X]^T [X]\right]^{-1} [X]^T Y \text{, де}$$

$$[X] = \begin{pmatrix} 1 & 3 & 7 \\ 1 & 4 & 9 \\ 1 & 5 & 11 \\ 1 & 7 & 12 \\ 1 & 9 & 12 \\ 1 & 11 & 15 \\ 1 & 10 & 18 \\ 1 & 12 & 21 \\ 1 & 11 & 22 \\ 1 & 12 & 24 \end{pmatrix} \qquad Y = \begin{pmatrix} 4 \\ 5 \\ 6 \\ 8 \\ 11 \\ 11 \\ 12 \\ 12 \\ 13 \\ 14 \end{pmatrix}$$

$$[X]^T[X]=\begin{pmatrix}1&1&1&1&1&1&1&1&1&1\\3&4&5&7&9&11&10&12&11&12\\7&9&11&12&12&15&18&21&22&24\end{pmatrix}\times\begin{pmatrix}1&3&7\\1&4&9\\1&5&11\\1&7&12\\1&9&12\\1&11&15\\1&10&18\\1&12&21\\1&11&22\\1&12&24\end{pmatrix}=\begin{pmatrix}10&84&151\\84&810&1431\\151&1431&2589\end{pmatrix}.$$

$$\left[[X]^T[X]\right]^{-1}=\begin{pmatrix}0{,}848978&-0{,}02401&-0{,}03625\\-0{,}02401&0{,}053163&-0{,}02798\\-0{,}03625&-0{,}02798&0{,}017968\end{pmatrix}.$$

$$[X]^T Y=\begin{pmatrix}1&1&1&1&1&1&1&1&1&1\\3&4&5&7&9&11&10&12&11&12\\7&9&11&12&12&15&18&21&22&24\end{pmatrix}\times\begin{pmatrix}4\\5\\6\\8\\11\\11\\12\\12\\13\\14\end{pmatrix}=\begin{pmatrix}96\\913\\1622\end{pmatrix}.$$

$$\hat{\beta}=\begin{pmatrix}0{,}848978&-0{,}02401&-0{,}03625\\-0{,}02401&0{,}053163&-0{,}02798\\-0{,}03625&-0{,}02798&0{,}017968\end{pmatrix}\times\begin{pmatrix}96\\913\\1622\end{pmatrix}=\begin{pmatrix}0{,}79\\0{,}84\\0{,}11\end{pmatrix}.$$

$$y=0{,}79+0{,}84x_1+0{,}11x_2\,.$$

$$\mathrm{VAR}(\hat{\beta})=\sigma_u^2\left[[X^T][X]\right]^{-1}=\begin{pmatrix}\sigma^2_{\hat{\beta}_1}&\sigma^2_{\hat{\beta}_1\hat{\beta}_2}&\sigma^2_{\hat{\beta}_1\hat{\beta}_3}\\\sigma^2_{\hat{\beta}_2\hat{\beta}_1}&\sigma^2_{\hat{\beta}_2}&\sigma^2_{\hat{\beta}_2\hat{\beta}_3}\\\sigma^2_{\hat{\beta}_3\hat{\beta}_1}&\sigma^2_{\hat{\beta}_3\hat{\beta}_2}&\sigma^2_{\hat{\beta}_3}\end{pmatrix},$$

where $\sigma_u^2=\dfrac{e'e}{n-k}$,

$\hat{y}_1=0{,}79+0{,}84\cdot 3+0{,}11\cdot 7=4{,}12$; $\hat{y}_2=0{,}79+0{,}84\cdot 4+0{,}11\cdot 9=5{,}19$; and so on $\hat{y}_{10}=0{,}79+0{,}84\cdot 12+0{,}11\cdot 24=13{,}65$.

e_1=4 -4,12 = -0,12; e_2 = 5 - 5,19 = -0,19; and so on e_{10}= 14 - 13,65 = 0,35.

$$e'\,e = \begin{pmatrix} -0,12 & -0,19 & -0,26 & -0,06 & 1,25 & -0,78 & 0,72 & -1,31 & 0,42 & 0,35 \end{pmatrix} \cdot \begin{pmatrix} -0,12 \\ -0,19 \\ -0,26 \\ -0,06 \\ 1,25 \\ -0,78 \\ 0,72 \\ -1,31 \\ 0,42 \\ 0,35 \end{pmatrix} = 4,82.$$

$\sigma_u^2 = 4,82/(10\text{-}3) = 0,69.$

$$\operatorname{var}\left(\hat{\beta}\right) = 0,69 \times \begin{pmatrix} 0,848978 & -0,02401 & -0,03625 \\ -0,02401 & 0,053163 & -0,02798 \\ -0,03625 & -0,02798 & 0,017968 \end{pmatrix} = \begin{pmatrix} 0,585795 & -0,01657 & -0,02501 \\ -0,01657 & 0,036682 & -0,01931 \\ -0,02501 & -0,01931 & 0,012398 \end{pmatrix}.$$

Topic. Multicollinearity

Task

Investigate the problem of multicollinearity for the next statistical sample

x_1	x_2	x_3
1,43	9,18	8,97
3,92	10,94	9,8
5,49	11,3	9,6
8,17	13,99	8,02
9,68	14,59	10,51
14,42	15,73	12,39
15,92	17,97	16,36
18,04	19,45	9,32
20,69	21,49	12,72
22,68	21,8	16,22

$$[R] = \begin{bmatrix} 1 & 0,99 & 0,70 \\ 0,99 & 1 & 0,66 \\ 0,70 & 0,66 & 1 \end{bmatrix}. \qquad \det[R] = 0,0077.$$

$$\chi_p{}^2 = -[n - 1 - \frac{1}{6}(2m+5)] \ln \det[R] = -\left[10 - 1 - \frac{1}{6}(6+5)\right] \ln 0,0077 = 34,85.$$

$\chi^2_{кр}(0,95;3) = 7,8.$

$$[Z]=[R]^{-1}=\begin{bmatrix} 72,48 & -67,98 & -5,83 \\ -67,98 & 65,54 & 4,29 \\ -5,83 & 4,29 & 2,25 \end{bmatrix}$$

$$t_{ij}=\frac{r_{ij}^{*}\sqrt{n-m-1}}{\sqrt{1-r_{ij}^{*2}}},$$

where $r_{ij}^{*}=\frac{-z_{ij}}{\sqrt{z_{ii}z_{jj}}}$; z_{ij}, z_{ii}, z_{jj} – elements of matrix $[Z]$.

$$r_{12}^{*}=\frac{-(-67,98)}{\sqrt{72,48\cdot 65,54}}=0,986; \quad t_{12}=\frac{0,986\cdot\sqrt{10-3-1}}{\sqrt{1-0,986^{2}}}=14,639;$$

$$r_{13}^{*}=\frac{-(-5,83)}{\sqrt{72,48\cdot 2,25}}=0,456; \quad t_{13}=\frac{0,456\cdot\sqrt{10-3-1}}{\sqrt{1-0,456^{2}}}=1,256;$$

$$r_{23}^{*}=\frac{-4,29}{\sqrt{65,54\cdot 2,25}}=-0,353; \quad t_{23}=\frac{-0,353\cdot\sqrt{10-3-1}}{\sqrt{1-0,353^{2}}}=-0,923.$$

k=*n*-*m*-1=10-3-1=6 and *p*=0,95 $t(0,95;6)=2,227$. For x_1 and x_2 $t_{12}>t(0,95;6)$. Multicollinearity exist for pair of factors x_1 and x_2 with significance *p*=0,95. Factor x_1 is deleted from the next analysis.

For x_2 and x_3: $[R]=\begin{bmatrix} 1 & 0,66 \\ 0,66 & 1 \end{bmatrix}$.

$[Z]=\begin{bmatrix} 1,785 & -1,184 \\ -1,184 & 1,785 \end{bmatrix}$, $\det[R]=0,560$.

Value $\chi_p^2=-\left[10-1-\frac{1}{6}(4+5)\right]\ln 0,560=4,346$ is lesser $\chi_{кр}^2(0,95;2)=6,0$.

Multicollinearity between x_2 and x_3 – absent.