Theme: Relationship of random variables in economics. Correlation dependence. Elements of regression analysis. Forecasting the characteristics of the foreign trade market

Part 2. Regression theory

Notions of a regression and a correlation are related. **A correlation analysis** estimates a strength of a stochastic correlation (зв’язки), **a regression analysis** investigates its form. Both types of analysis are used for explanations of cause-and-effect relations between phenomenons and for defining a presence or an absence of a correlation.

Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the "statistical significance" of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship.

Let *Y* denote the "dependent" variable whose values you wish to predict.

let ***X*** denote the “independent” variable from which you wish to predict it

There are such types ***a correlation and a regression***:

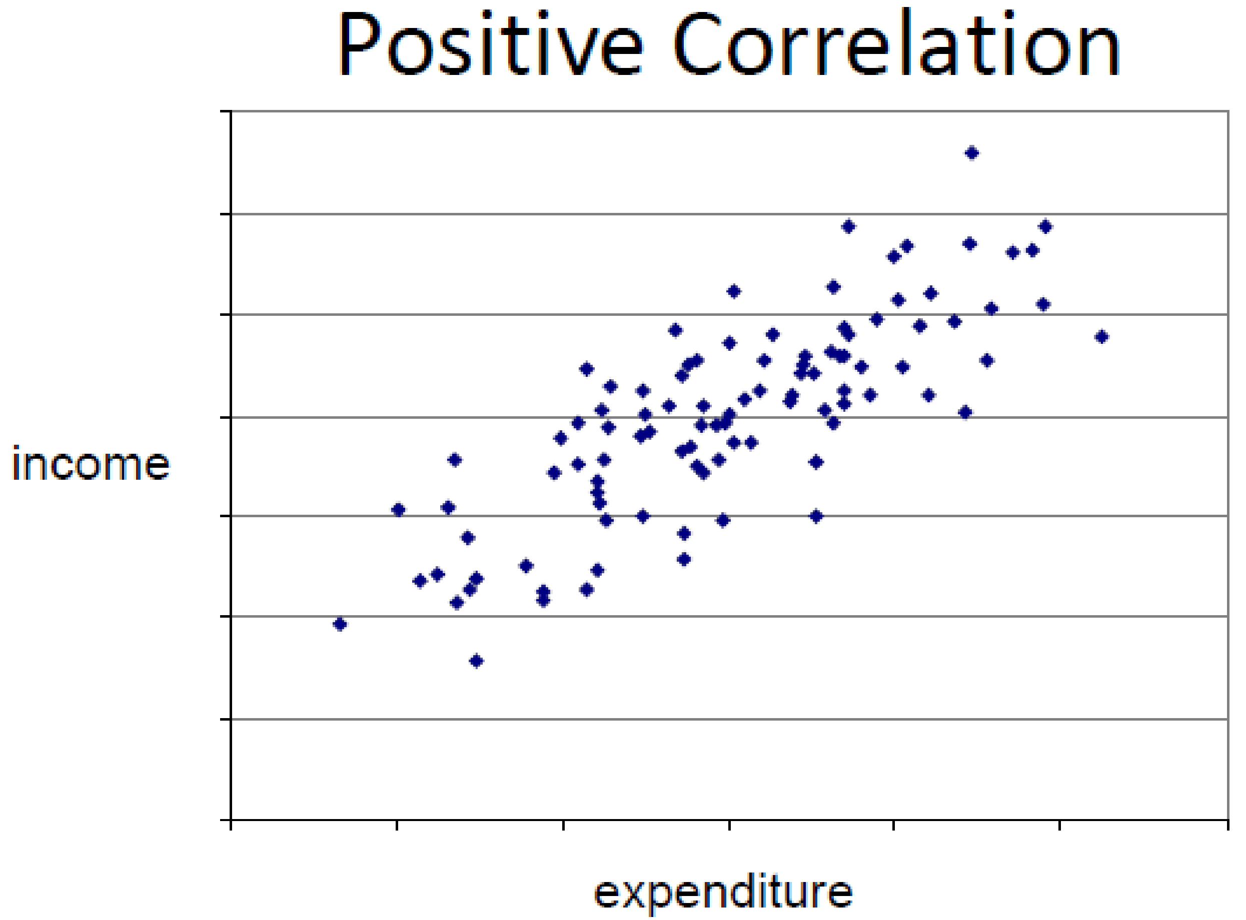
1)relative to a character of a correlation and regression we have a positive or negative correlation and regression. A positive value means that an increase (decrease) of an argument $\boldsymbol{x}$ is related to an increase (decrease) of a functional factor $\boldsymbol{y}$. A positive value means that an increase (decrease) of an argument $\boldsymbol{x}$ is related to a decrease (increase) of a functional factor $\boldsymbol{y}$.

Example

Correlation

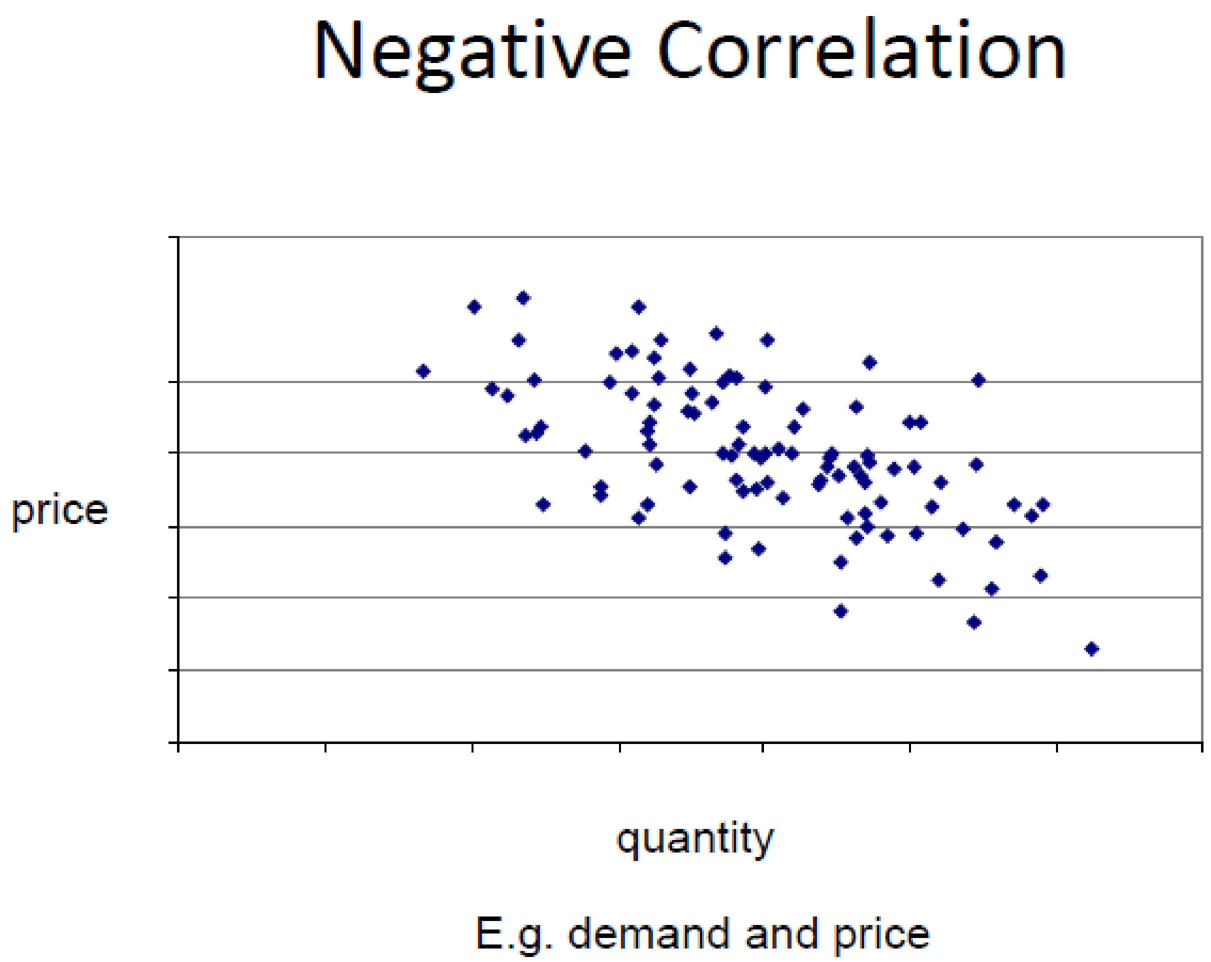
- Correlation examines the relationships between pairs of variables, for example
  - between the price of doughnuts and the demand for them
  - between economic growth and life expectancy
  - between hair colour and hourly wage
  - between rankings
- Such analyses can be useful for formulating policies

Example



E.g. income and food expenditure

Example



There are such types ***a correlation and a regression***:

2) ***relative to a number of variables***: a pair or multiple correlation and regression.

Example

A pair regression is $\overline{y} = f(x)$

Example

A pair regression is $\overline{y} = f(x)$

A multiple regression is $\overline{y} = f(x_1, x_2, x_3, x_4)$

There are such types ***a correlation and a regression***:

3) ***relative to a correlation form***: a linear or nonlinear correlation and regression.

BASIC TYPES OF EQUATIONS

Linear equation:

$$\tilde{y}_x = b_0 + b_1 x$$

Power function:

$$\tilde{y}_x = b_0 \cdot x^{b_1}$$

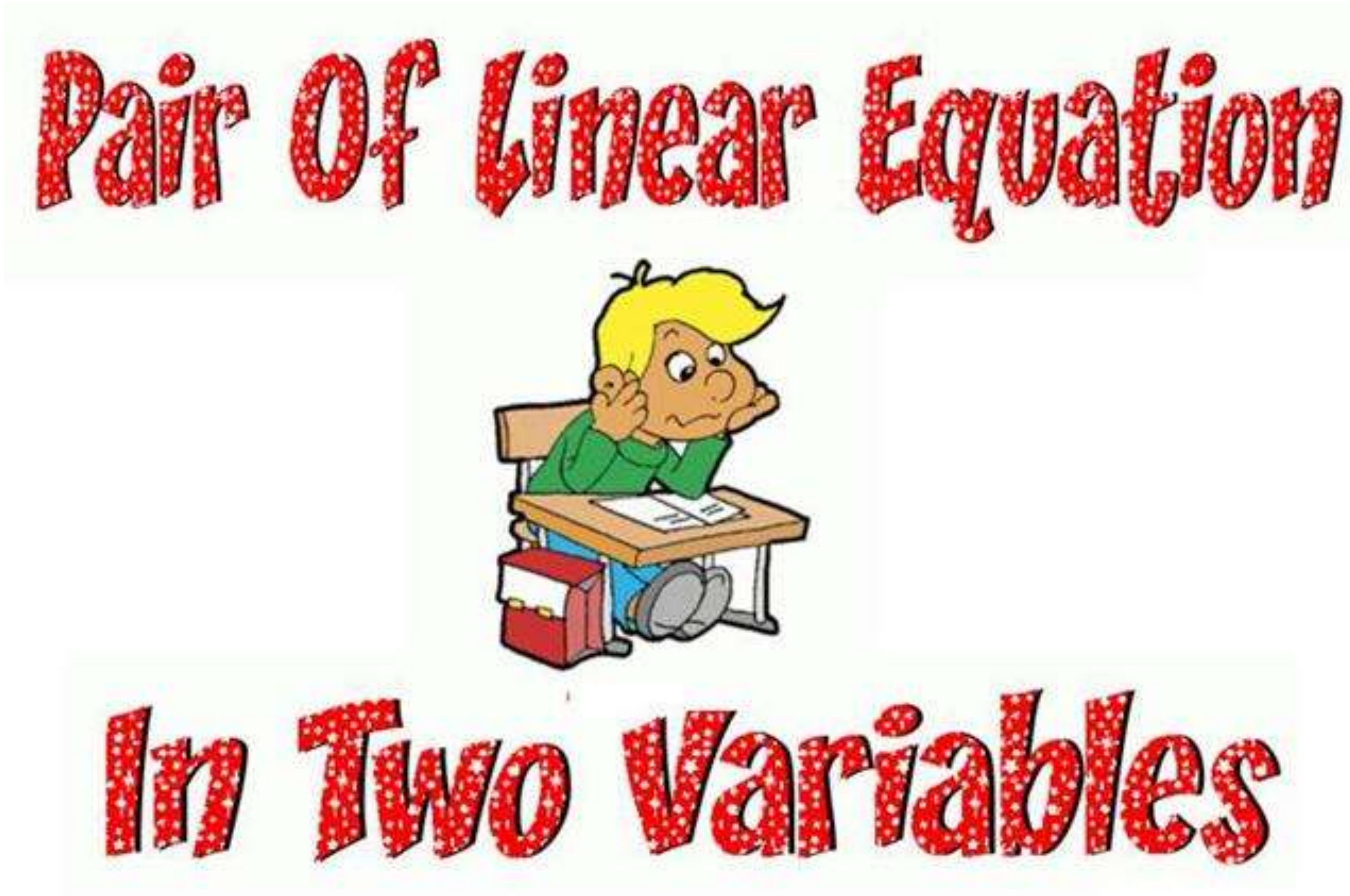
Exponential function:

$$\tilde{y}_x = b_0 \cdot b_1^x$$

Hyperbolic function:

$$\tilde{y}_x = \frac{1}{b_0 + b_1 \cdot x}$$

Linear regression analysis is the most widely used of all statistical techniques: it is the study of *linear*, *additive* relationships between variables.


Pair Of Linear Equation
In Two Variables

A pair regression is an equation of two variables Y and X.

The equation of a linear regression

$$\tilde{y}_x = b_0 + b_1 x$$

where b_0 and b_1 are parameters of the equation.

The equation $\overline{y}_x = f(x)$ is called **the regression equation *of Y upon X***, and the function $f(x)$ **the regression of *Y upon X***, and its graph is ***the regression line of Y upon X***.

We can use the other equation: $\overline{x}_y = \varphi(y)$. It is **the regression equation *of X upon Y.***

LEAST SQUARES METHOD

The condition of a minimization of the given function is reduced to a solving a system of equations. Results of this system are

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \qquad b_0 = \overline{y} - b_1 \cdot \overline{x}$$

The equation of a linear regression

Results of this system are

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \qquad b_0 = \overline{y} - b_1 \cdot \overline{x}$$

where

$$\mu_{xy} = \overline{xy} - \overline{x}\,\overline{y} = \frac{1}{n}\sum_1^n x_i y_i - \overline{x}\,\overline{y}$$

is called the coefficient of the covariation.

The equation of a linear regression

Results of this system are

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \qquad b_0 = \overline{y} - b_1 \cdot \overline{x}$$

where

$$\sigma_x^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{n}\sum_1^n x_i^2 - \overline{x}^2$$

is called the variance of *x*

The equation of a linear regression

Results of this system are

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} \qquad b_0 = \bar{y} - b_1 \cdot \bar{x}$$

where

$$\bar{x} = \frac{1}{n}\sum_1^n x_i$$

is called the mean value of *x*

The equation of a linear regression

Results of this system are

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \qquad b_0 = \overline{y} - b_1 \cdot \overline{x}$$

where

$$\overline{y} = \frac{1}{n} \sum_1^n y_i$$

is called the mean value of *y*

Example

The dependence between variables X and Y is obtained on the base of the experiment and it is given as the following table:

x_i	1	1,5	2	2,5	3
y_i	2,15	2,3	2,6	2,8	2,5

Supposing the linear dependence $\hat{y}_x = b_0 + b_1 x$ between ***X*** and ***Y***

:

Example

Let's fill in this table:

No	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
1	1	2,15	2,15	1	4,6225
2	1,5	2,3	3,45	2,25	5,29
3	2	2,6	5,2	4	6,76
4	2,5	2,8	7	6,25	7,84
5	3	2,5	7,5	9	6,25
Σ	10	12,35	25,3	22,5	30,7625

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

.

:

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

$$\sigma_x^2 = \overline{x^2} - \left(\bar{x}\right)^2 = 4{,}5 - 2^2 = 0{,}5$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

$$\sigma_x^2 = \overline{x^2} - \left(\bar{x}\right)^2 = 4{,}5 - 2^2 = 0{,}5$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{0{,}5} = 0{,}71$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12,35}{5} = 2,47$$

$$\overline{x^2} = \frac{22,5}{5} = 4,5$$

$$\overline{y^2} = \frac{30,7625}{5} = 6,1525$$

$$\overline{xy} = \frac{25,3}{5} = 5,06$$

$$\sigma_y^2 = \overline{y^2} - \left(\bar{y}\right)^2 = 6,1525 - 2,47^2 = 0,0516$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12,35}{5} = 2,47$$

$$\overline{x^2} = \frac{22,5}{5} = 4,5$$

$$\overline{y^2} = \frac{30,7625}{5} = 6,1525$$

$$\overline{xy} = \frac{25,3}{5} = 5,06$$

.

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{0,0516} \approx 0,2272$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2 \qquad \bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5 \qquad \overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

$$\mu_{xy} = \overline{xy} - \bar{x} \cdot \bar{y} = 5{,}06 - 2 \cdot 2{,}47 = 0{,}12$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12,35}{5} = 2,47$$

$$\overline{x^2} = \frac{22,5}{5} = 4,5$$

$$\overline{y^2} = \frac{30,7625}{5} = 6,1525$$

$$\overline{xy} = \frac{25,3}{5} = 5,06$$

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{0,12}{0,5} = 0,24$$

.

:

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2$$

$$\bar{y} = \frac{12{,}35}{5} = 2{,}47$$

$$\overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06$$

$$b_1 = \frac{\mu_{xy}}{\sigma_x^2} = \frac{0{,}12}{0{,}5} = 0{,}24$$

.

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 2{,}47 - 0{,}24 \cdot 2 = 1{,}99$$

Example

Let's calculate:

$$\bar{x} = \frac{10}{5} = 2 \qquad \bar{y} = \frac{12,35}{5} = 2,47 \qquad \overline{x^2} = \frac{22,5}{5} = 4,5$$

$$\overline{xy} = \frac{25,3}{5} = 5,06$$

The sought equation has the form:

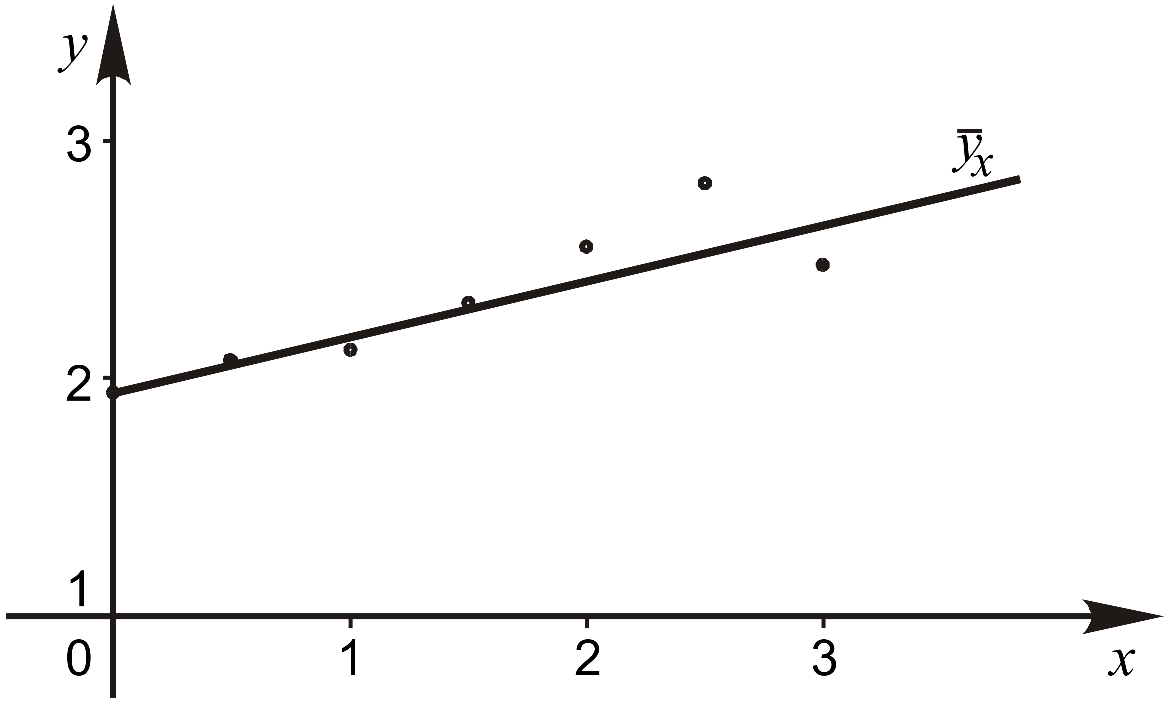
$$\tilde{y}_x = b_0 + b_1 x \qquad \tilde{y}_x = 1,99 + 0,24x$$

.

.

Example

Let’s mark the empirical points and plot the sought theoretical regression line $\hat{y}_x$



.

:

Example 2

Let's check this equation:

$$\overline{y}_{x=1,5} = 0,24 \cdot 1,5 + 1,99 = 2,35$$

$$\overline{y}_{x=2,5} = 0,24 \cdot 2,5 + 1,99 = 2,59$$

Deviations: $\varepsilon_1 = \left|2,35 - 2,3\right| = 0,05$

$$\varepsilon_2 = \left|2,59 - 2,6\right| = 0,01$$

.

:

Example

The theoretical regression equation is

$$\tilde{y}_x = 1{,}99 + 0{,}24x$$

Explanation: the coefficient $b_1 = 0{,}24$ shows the increasing X by 1 unit gives the increasing Y by 0,24 units.

A sign of a regression coefficient and a sign of a covariation coincide: the sign of b_1 is «+» then there is a positive regression, the sign «–» means a negative regression.

A correlation coefficient, its properties and explanation

Since a correlation moment (a covariation) μ_{xy} is a dimensional characteristic, then one introduces an additional nondimensional characteristic (***a correlation coefficient***) as a coefficient of a **strength** (тіснота) of a linear correlation:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \sigma_y}$$

Properties of a correlation coefficient:

1) If random variable X and Y are linearly independent, then $r_{xy} = 0$ (since $\mu_{xy} = 0$).

2) random variable X and Y are linearly dependent, then $r_{xy} = 1$.

3) $\left| r_{xy} \right| \leq 1$, $-1 \leq r \leq 1$

4) $r_{xy} = r_{yx}$ according to the definition

Example

We get:

$$\bar{x} = \frac{10}{5} = 2 \qquad \bar{y} = \frac{12{,}35}{5} = 2{,}47 \qquad \overline{x^2} = \frac{22{,}5}{5} = 4{,}5$$

$$\overline{xy} = \frac{25{,}3}{5} = 5{,}06 \qquad \overline{y^2} = \frac{30{,}7625}{5} = 6{,}1525$$

Example 4

We get

$$\mu_{xy} = \overline{xy} - \overline{x} \cdot \overline{y} = 5{,}06 - 2 \cdot 2{,}47 = 0{,}12$$

$$\sigma_x = \sqrt{\sigma_x^2} = \sqrt{0{,}5} = 0{,}71$$

$$\sigma_y = \sqrt{\sigma_y^2} = \sqrt{0{,}0516} \approx 0{,}2272$$

The correlation coefficient:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y} = \frac{0{,}12}{0{,}71 \cdot 0{,}2272} = 0{,}7471$$

- The value of r is such that $-1 \le r \le +1$. The + and – signs are used for positive linear correlations and negative linear correlations, respectively.
- *Positive correlation:* If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values indicate a relationship between x and y variables such that as values for x increases, values for y also increase.
- *Negative correlation:* If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.
- *No correlation:* If there is no linear correlation or a weak linear correlation, r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables
- Note that r is a dimensionless quantity; that is, it does not depend on the units employed.
- A *perfect* correlation of ± 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, the slope of this line is negative.
- A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the "type" of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Values of *r*

We will use this explanation:

<u>*Conclusion*:</u> If $|r| < 0{,}35$ then this correlation is **weak**, if $0{,}35 \le |r| \le 0{,}7$ then this correlation is **moderate**, if $|r| > 0{,}7$ then this correlation is **strong**.

Example

The correlation coefficient:

$$r_{xy} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y} = \frac{0{,}12}{0{,}71 \cdot 0{,}2272} = 0{,}7471$$

Since this coefficient ***r*** is greater than 0.7, then this linear correlation is strong (сильна кореляція).

The mean error of the approximation

The limit of the mean error will be allowable (допустима помилка) if it equals 8-10%.

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100\%$$

EXAMPLE

№	y_i	$\tilde{y}_i$	$y_i - \tilde{y}_i$	$\frac{y_i - \tilde{y}_i}{y_i}$	$\left\lvert\frac{y_i - \tilde{y}_i}{y_i}\right\rvert$
1	2,15	2,23	-0,08	-0,0372	0,0372
2	2,3	2,35	-0,05	-0,0217	0,0217
3	2,6	2,47	0,13	0,05	0,05
4	2,8	2,59	0,21	0,075	0,075
5	2,5	2,71	-0,21	-0,084	0,084
Σ	12,35				0,2679

EXAMPLE

The mean error equals

$$\overline{A} = \frac{1}{5} \cdot 0{,}2679 \cdot 100\% \approx 0{,}0536 \cdot 100\% = 5{,}36\%$$

It's the allowable limit.

A determination coefficient

The **coefficient of determination**, denoted R^2 or r^2 and pronounced **R squared**, is a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variable

Coefficient of Determination, r^2 or R^2:

- The *coefficient of determination*, r^2, is useful because it gives the proportion of the variance (fluctuation) of one variable that is predictable from the other variable. It is a measure that allows us to determine how certain one can be in making predictions from a certain model/graph.
- The *coefficient of determination* is the ratio of the explained variation to the total variation.
- The *coefficient of determination* is such that $0 \leq r^2 \leq 1$, and denotes the strength of the linear association between x and y.
- The *coefficient of determination* represents the percent of the data that is the closest to the line of best fit. For example, if $r = 0.922$, then $r^2 = 0.850$, which means that 85% of the total variation in y can be explained by the linear relationship between x and y (as described by the regression equation). The other 15% of the total variation in y remains unexplained.
- The *coefficient of determination* is a measure of how well the regression line represents the data. If the regression line passes exactly through every point on the scatter plot, it would be able to explain all of the variation. The further the line is away from the points, the less it is able to explain.

EXAMPLE

Let’s continue to solve **EXAMPLE**

$$r_{xy} = 0{,}7471$$

$$R^2 = r_{xy}^2 = 0{,}7471^2 = 0{,}5582$$

It means that 55,82% of the total variation in ***y*** can be explained by the linear relationship between ***x*** and ***y*** (as described by the regression equation). The other 44,18% of the total variation in ***y*** remains unexplained.

A statistical method that explains how much of the variability of a factor can be caused or explained by its relationship to another factor.

Coefficient of determination is used in trend analysis. It is computed as a value between 0 (0 percent) and 1 (100 percent). The higher the value, the better the fit. Coefficient of determination is symbolized by r^2 because it is square of the coefficient of correlation symbolized by r. The coefficient of determination is an important tool in determining the degree of linear-correlation of variables ('goodness of fit') in regression analysis.

Coefficient of Determination

The coefficient of determination is ...

- the percent of the variation that can be explained by the regression equation.
- the explained variation divided by the total variation
- the square of r

What's all this variation stuff?

Every sample has some variation in it (unless all the values are identical, and that's unlikely to happen). The total variation is made up of two parts, the part that can be explained by the regression equation and the part that can't be explained by the regression equation.

$$\sum (y - \bar{y})^2 = \sum (y' - \bar{y})^2 + \sum (y - y')^2$$

$$\textit{total} = \textit{explained} + \textit{unexplained}$$

Well, the ratio of the explained variation to the total variation is a measure of how good the regression line is. If the regression line passed through every point on the scatter plot exactly, it would be able to explain all of the variation. The further the line is from the points, the less it is able to explain.

F-test (Fisher test) for verification of
a statistical significance
(статистична значущість)
and
a reliability (надійність)

Let's make **the assumption (hypothesis H_0)** about a statistical nonsignificance (insignificance) of an equation and an indicator of a correlation strength (гіпотеза про статистичну незначущість рівняння і показника тісноти зв'язку).

We compare the tabular value F_{tabl} for the corresponding significance level with the empirical value:

$$F_{emp} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2)$$

We compare the tabular value

$$F_{tabl}\left(k_1; k_2\right)$$

for the corresponding significance level with the empirical value:

$$F_{emp} = \frac{r_{xy}^2}{1 - r_{xy}^2}\left(n - 2\right)$$

We compare the tabular value

$$F_{tabl}(1; n-2)$$

for the corresponding significance level with the empirical value:

$$F_{emp} = \frac{r_{xy}^2}{1-r_{xy}^2}(n-2)$$

If $F_{tabl} < F_{emp}$

then the hypothesis H_0 is rejected, a statistical significance and a reliability of a regression equation is accepted.

If $F_{tabl} > F_{emp}$

then the hypothesis H_0 is accepted, a statistical nonsignificance of an equation is accepted.

$\alpha = 0{,}05$

k_2	k_1											
	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	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4
3	10,1	9,55	9,28	9,13	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	4,16	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,!0	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

$$\alpha = 0{,}05$$

k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
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
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35	2,31	2,28
24	4,26	3,40	3,01	2,78	2,62	2,51	2,43	2,36	2,30	2,26	2,22	2,18
30	4,!7	3,32	2,92	2,69	2,53	2,42	2,34	2,27	2,21	2,16	2,!2	2,09
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,!8	2,12	2,07	2,04	2,00
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,02	1,98	1,95
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,01	1,97	1,93	1,89
100	3,94	3,09	2,70	2,46	2,30	2,19	2,10	2,03	1,97	1,92	1,88	1,85
200	3,89	3,04	2,65	2,41	2,26	2,14	2,05	1,98	1,92	1,87	1,83	1,80
400	3,86	3,02	2,62	2,39	2,23	2,12	2,03	1,96	1,90	1,85	1,81	1,78
∞	3,84	2,99	2,60	2,37	2,21	1,09	2,01	1,94	1,88	1,83	1,79	1,75

$\alpha = 0,01$

k_2	k_1									
	1	2	3	4	5	6	8	12	24	∞
1	4052	4999	5403	5625	5764	5859	5981	6106	6234	6366
2	98,5	99,0	99,2	99,3	99,3	99,4	99,3	99,4	99,5	99,5
3	34,1	30,8	29,5	28,7	28,2	27,9	27,5	27,1	26,6	26,1
4	21,2	18,0	16,7	16,0	15,5	15,2	14,8	14,4	13,9	13,5
5	16,3	13,3	12,1	11,4	11,0	10,7	10,3	9,9	9,5	9,6
6	13,7	10,9	9,8	9,2	8,8	8,5	8,1	7,7	7,3	6,9
7	12,3	9,6	8,5	7,9	7,5	7,2	6,8	6,5	6,1	5,7
8	11,3	8,7	7,6	7,0	6,6	6,5	6,0	5,7	5,3	4,9
9	10,6	8,0	7,0	6,4	6,1	5,8	5,5	5,1	4,7	4,3
10	10,0	7,6	6,6	6,0	5,6	5,4	5,1	4,7	4,3	3,9
11	9,7	7,2	6,2	5,7	5,3	5,1	4,7	4,4	4,0	3,6
12	9,3	6,9	6,0	5,4	5,1	4,8	4,5	4,2	3,8	3,4
13	9,1	6,7	5,7	5,2	4,9	4,6	4,3	4,0	3,6	3,2

$$\alpha = 0{,}01$$

k_2	k_1									
	1	2	3	4	5	6	8	12	24	∞
14	8,9	6,5	5,6	5,0	4,7	4,5	4,1	3,8	3,4	3,0
15	8,7	6,4	5,4	4,9	4,6	4,3	4,0	3,7	3,3	2,9
16	8,5	6,2	5,3	4,8	4,4	4,2	3,9	3,6	3,2	2,8
17	8,4	6,1	5,2	4,7	4,3	4,1	3,8	3,5	3,1	2,7
18	8,3	6,0	5,1	4,6	4,3	4,0	3,7	3,4	3,0	2,6
19	8,2	5,9	5,0	4,5	4,2	3,9	3,6	3,3	2,9	2,4
20	8,1	5,9	4,9	4,4	4,1	3,9	3,6	3,2	2,9	2,4
22	7,9	5,7	4,8	4,3	4,0	3,8	3,5	3,1	2,8	2,3
24	7,8	5,6	4,7	4,2	3,9	3,7	3,3	3,0	2,7	2,2
26	7,7	5,5	4,6	4,1	3,8	3,6	3,3	3,0	2,6	2,1
28	7,6	5,5	4,6	4,1	3,8	3,5	3,2	2,9	2,5	2,1
30	7,6	5,4	4,5	4,0	3,7	3,5	3,2	2,8	2,5	2,0
40	7,3	5,2	4,3	3,8	3.5	3,3	3,0	2,7	2,3	1,8
60	7,1	5,0	4,1	3,7	3,3	3,1	2,8	2,5	2,1	1,6
120	6,9	4,8	4,0	3,5	3,2	3,0	2,7	2,3	2,0	1,4
∞	6,6	4,6	3,8	3,3	3,0	2,8	2,5	2,2	1,8	1,0

Example

Let’s calculate:

$$F_{emp} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2)$$

Example

Let's calculate:

$$F_{emp} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n-2)$$

$$F_{emp} = \frac{0{,}7471^2}{1 - 0{,}7471^2}(5-2) \approx 3{,}79$$

Example

Let’s calculate:

$$F_{emp} = \frac{0,7471^2}{1 - 0,7471^2}(5-2) \approx 3,79$$

Example

$$F_{emp} = \frac{0,7471^2}{1-0,7471^2}(5-2) \approx 3,79$$

Let’s compare with the tabular value:

$$F_{tabl}(k_1; k_2) = F_{tabl}(1; n-2) = F_{tabl}(1; 5-2) = F_{tabl}(1; 3)$$

Example

$$F_{emp} = \frac{0,7471^2}{1-0,7471^2}(5-2) \approx 3,79$$

Let’s compare with the tabular value:

$$F_{tabl}(k_1; k_2) = F_{tabl}(1; n-2) = F_{tabl}(1; 3) = 10,1$$

Example

$$F_{emp} = \frac{0,7471^2}{1-0,7471^2}(5-2) \approx 3,79$$

Let’s compare with the tabular value:

$$F_{tabl}(k_1; k_2) = F_{tabl}(1; 3) = 10,1$$

We get:

$$F_{tabl} > F_{emp}$$

Example

Let's compare with the tabular value:

$$F_{tabl} > F_{emp}$$

$$10{,}1 > 3{,}79$$

Conclusion:

Example

Let's compare with the tabular value:

$$F_{tabl} > F_{emp}$$

$$10{,}1 > 3{,}79$$

Conclusion: the hypothesis H_0 is accepted, i.e. the statistical nonsignificance of this regression equation is accepted with 95 %.

t-test (Student test) for verification of

a significance of

a regression equation and

a correlation

Let’s make **the assumption (hypothesis H_0)** about a random character of model parameters or correlation coefficients.

We compare the tabular value t_{tabl} for the corresponding significance level with the empirical values:

$$t_{b_i} = \frac{b_i}{m_{b_i}}$$

where m_{b_i} are mean errors of a linear regression.

We compare the tabular value t_{tabl} for the corresponding significance level with the empirical values:

$$t_{b_0} = \frac{b_0}{m_{b_0}}$$

$$m_{b_0} = \sqrt{\frac{\sum_i \left(y_i - \tilde{y}_i\right)^2}{n-2} \frac{\sum_i x_i^2}{n\sum_i \left(x_i - \bar{x}\right)^2}} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2}$$

We compare the tabular value t_{tabl} for the corresponding significance level with the empirical values:

$$t_{b_1} = \frac{b_1}{m_{b_1}}$$

$$m_{b_1} = \sqrt{\frac{\sum_i (y_i - \tilde{y}_i)^2 \Big/ (n-2)}{\sum_i (x_i - \bar{x})^2}} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}}$$

We compare the tabular value t_{tabl} for the corresponding significance level with the empirical values:

$$t_r = \frac{r}{m_r}$$

where m_r is the mean error of a correlation coefficient.

We compare the tabular value t_{tabl} for the corresponding significance level with the empirical values:

$$t_r = \frac{r}{m_r}$$

$$m_r = \sqrt{\frac{1-r_{xy}^2}{n-2}}$$

If $t_{tabl} < t_{emp}$

then the hypothesis H_0 is rejected, a statistical significance of model coefficients is accepted.

If $t_{tabl} > t_{emp}$

then the hypothesis H_0 is accepted, a statistical nonsignificance of model coefficients and a randomness of their forming are accepted.

Student's distribution

α				
k	$\alpha = 0{,}01$	$\alpha = 0{,}02$	$\alpha = 0{,}05$	$\alpha = 0{,}1$
1	63,7	31,82	12,7	6,31
2	9,92	6,97	4,30	2,92
3	5,84	4,54	3,18	2,35
4	4,60	3,75	2,78	2,13
5	4,03	3,37	2,57	2,01
6	3,71	3,14	2,45	1,94
7	3,50	3,00	2,36	1,89
8	3,36	2,90	2,31	1,86
9	3,25	2,82	2,26	1,83
10	3,17	2,76	2,23	1,81
11	3,11	2,72	2,20	1,80
12	3,05	2,68	2,18	1,78
13	3,01	2,65	2,16	1,77
14	2,98	2,62	2,14	1,76

Student's distribution

α				
k	$\alpha = 0{,}01$	$\alpha = 0{,}02$	$\alpha = 0{,}05$	$\alpha = 0{,}1$
14	2,98	2,62	2,14	1,76
15	2,95	2,60	2,13	1,75
16	2,92	2,58	2,12	1,75
17	2,90	2,57	2,11	1,74
18	2,88	2,55	2,10	1,73
19	2,86	2,54	2,09	1,73
20	2,85	2,53	2,09	1,73
21	2,83	2,52	2,08	1,72
22	2,82	2,51	2,07	1,72
23	2,81	2,50	2,07	1,71
24	2,80	2,49	2,06	1,71
25	2,79	2,49	2,06	1,71
26	2,78	2,48	2,06	1,71
27	2,77	2,47	2,05	1,71
28	2,76	2,46	2,05	1,70
30	2,75	2,46	2,04	1,70
40	2,70	2,42	2,02	1,68
60	2,66	2,39	2,00	1,67
120	2,62	2,36	1,98	1,66

EXAMPLE

№	y_i	$\tilde{y}_i$	$y_i - \tilde{y}_i$	$(y_i - \tilde{y}_i)^2$
1	2,15	2,23	-0,08	0,0064
2	2,3	2,35	-0,05	0,0025
3	2,6	2,47	0,13	0,0169
4	2,8	2,59	0,21	0,0441
5	2,5	2,71	-0,21	0,0441
Σ	12,35			0,114

EXAMPLE

The variance of residuals (залишкова дисперсія):

$$\sigma^2_{residuals} = \frac{\sum_i (y_i - \tilde{y}_i)^2}{n-2}$$

EXAMPLE

The variance of residuals (залишкова дисперсія):

$$\sigma^2_{residuals} = \frac{\sum_i \left(y_i - \tilde{y}_i\right)^2}{n-2} = \frac{0{,}114}{5-2} = 0{,}038$$

EXAMPLE

The variance of residuals (залишкова дисперсія):

$$\sigma^2_{residuals} = \frac{\sum_i (y_i - \tilde{y}_i)^2}{n-2} = \frac{0,114}{5-2} = 0,038$$

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0,038} = 0,195$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^{2}_{residuals}} = \sqrt{0,038} = 0,195$$

$$m_{b_1} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}} =$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_1} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}} = \frac{0{,}195}{0{,}71 \cdot \sqrt{5}} \approx 0{,}123$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_1} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}} = \frac{0{,}195}{0{,}71 \cdot \sqrt{5}} \approx 0{,}123$$

$$t_{b_1} = \frac{b_1}{m_{b_1}} =$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_1} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}} = \frac{0{,}195}{0{,}71 \cdot \sqrt{5}} \approx 0{,}123$$

$$t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{0{,}24}{0{,}123} =$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_1} = \frac{\sigma_{residuals}}{\sigma_x \sqrt{n}} = \frac{0{,}195}{0{,}71 \cdot \sqrt{5}} \approx 0{,}123$$

$$t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{0{,}24}{0{,}123} = 1{,}95$$

EXAMPLE (conclusions)

t-test for b_0

$$t_{0,05}(n-2) = t_{0,05}(5-2) = t_{0,05}(3) = 3{,}18$$

EXAMPLE (conclusions)

t-test for b_0

$$t_{0,05}(n-2) = t_{0,05}(5-2) = t_{0,05}(3) = 3{,}18$$

$$t_{0,05}(3) = 3{,}18 > t_{b1} = 1{,}95$$

Hypothesis H_0 is accepted, i.e. the statistical nonsignificance of the coefficient b_1 is accepted with 95%.

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_0} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2} =$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0,038} = 0,195$$

$$m_{b_0} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2} = 0,195 \cdot \frac{1}{5 \cdot 0,71} \cdot \sqrt{22,5} \approx 0,261$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_0} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2} = 0{,}195 \cdot \frac{1}{5 \cdot 0{,}71} \cdot \sqrt{22{,}5} \approx 0{,}261$$

$$t_{b_0} = \frac{b_0}{m_{b_0}}$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_0} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2} = 0{,}195 \cdot \frac{1}{5 \cdot 0{,}71} \cdot \sqrt{22{,}5} \approx 0{,}261$$

$$t_{b_0} = \frac{b_0}{m_{b_0}} = \frac{1{,}99}{0{,}261} =$$

EXAMPLE

The mean errors of parameters of the linear regression:

$$\sigma_{residuals} = \sqrt{\sigma^2_{residuals}} = \sqrt{0{,}038} = 0{,}195$$

$$m_{b_0} = \sigma_{residuals} \frac{1}{n\sigma_x} \sqrt{\sum_i x_i^2} = 0{,}195 \cdot \frac{1}{5 \cdot 0{,}71} \cdot \sqrt{22{,}5} \approx 0{,}261$$

$$t_{b_0} = \frac{b_0}{m_{b_0}} = \frac{1{,}99}{0{,}261} = 7{,}624$$

EXAMPLE (conclusions)

t-test for b_1

$$t_{0,05}(n-2)=t_{0,05}(5-2)=t_{0,05}(3)=3{,}18$$

EXAMPLE (conclusions)

t-test for b_1

$$t_{0,05}(n-2) = t_{0,05}(5-2) = t_{0,05}(3) = 3{,}18$$

$$t_{0,05}(3) = 3{,}18 < t_{b_0} = 7{,}624$$

Hypothesis H_0 is rejected, i.e. the statistical significance of the coefficient b_0 is accepted with 95%.

EXAMPLE

The mean error of the correlation coefficient:

$$r_{xy} = 0,7471$$

EXAMPLE

The mean error of the correlation coefficient:

$$r_{xy} = 0{,}7471$$

$$m_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}}$$

EXAMPLE

The mean error of the correlation coefficient:

$$r_{xy} = 0,7471$$

$$m_r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0,7471^2}{5 - 2}} \approx 0,384$$

EXAMPLE

The mean error of the correlation coefficient:

$$r_{xy} = 0{,}7471$$

$$m_r = \sqrt{\frac{1 - r_{xy}^2}{n - 2}} = \sqrt{\frac{1 - 0{,}7471^2}{5 - 2}} \approx 0{,}384$$

$$t_r = \frac{r_{xy}}{m_r} = \frac{0{,}7471}{0{,}384} \approx 1{,}95$$

EXAMPLE (conclusions)

t-test for r

$$t_{0,05}(n-2)=t_{0,05}(5-2)=t_{0,05}(3)=3,18$$

EXAMPLE (conclusions)

t-test for r

$$t_{0,05}(n-2) = t_{0,05}(5-2) = t_{0,05}(3) = 3{,}18$$

$$t_{0,05}(3) = 3{,}18 > t_r = 1{,}95$$

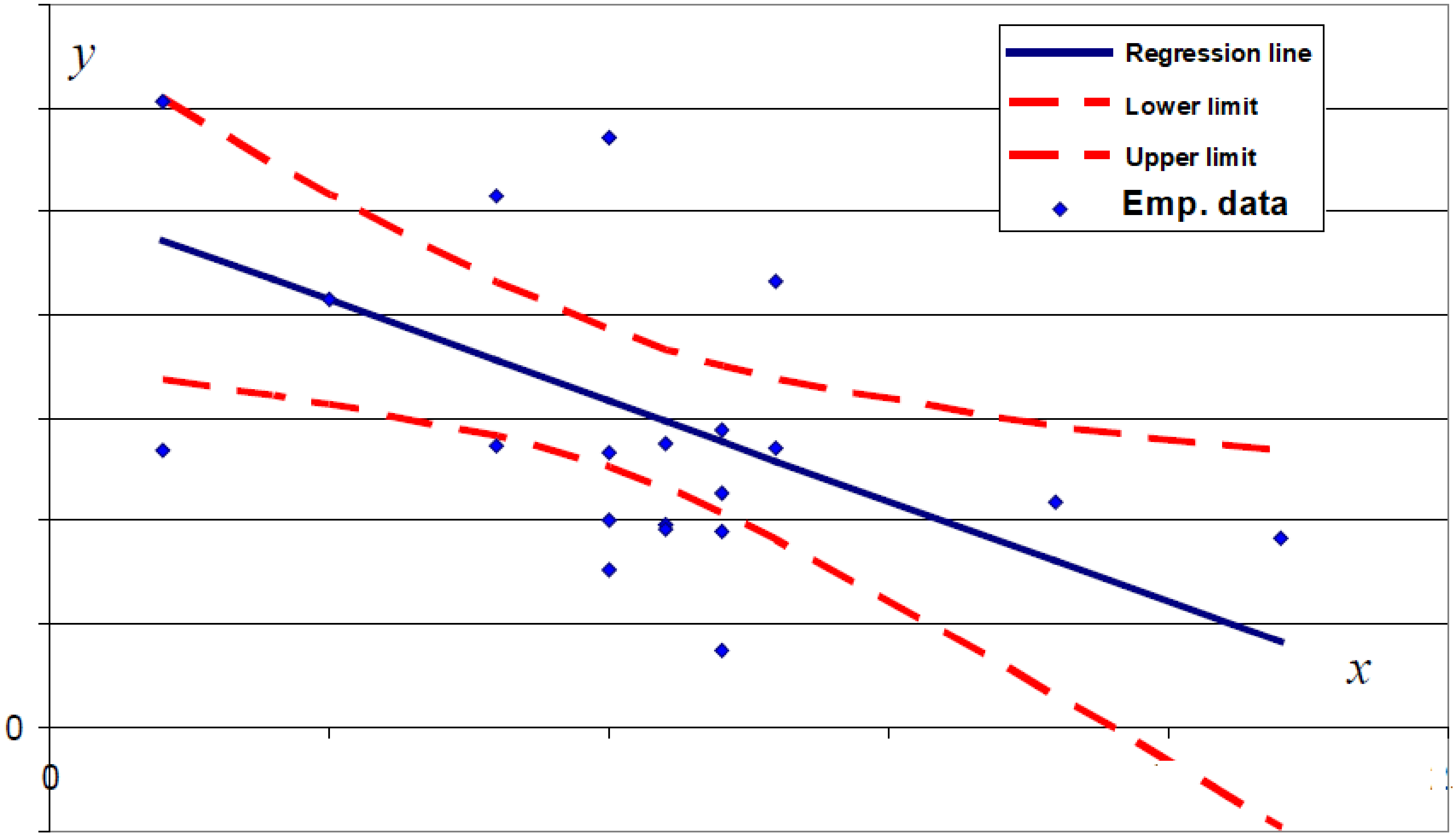
Hypothesis H_0 is accepted, i.e. the statistical nonsignificance of the correlation coefficient is accepted with probability 95%.

EXAMPLE (the built-in function LINEST

b1	0,24	1,99	b0
m b1	0,12328828	0,261533937	m b0
R^2	0,558139535	0,194935887	Sigma res
Femp	3,789473684	3	n-2
SSR	0,144	0,114	SSE

Confidence interval with the confidence probability γ or significance level α

Confidence interval with the confidence probability γ



Confidence interval with the confidence probability γ

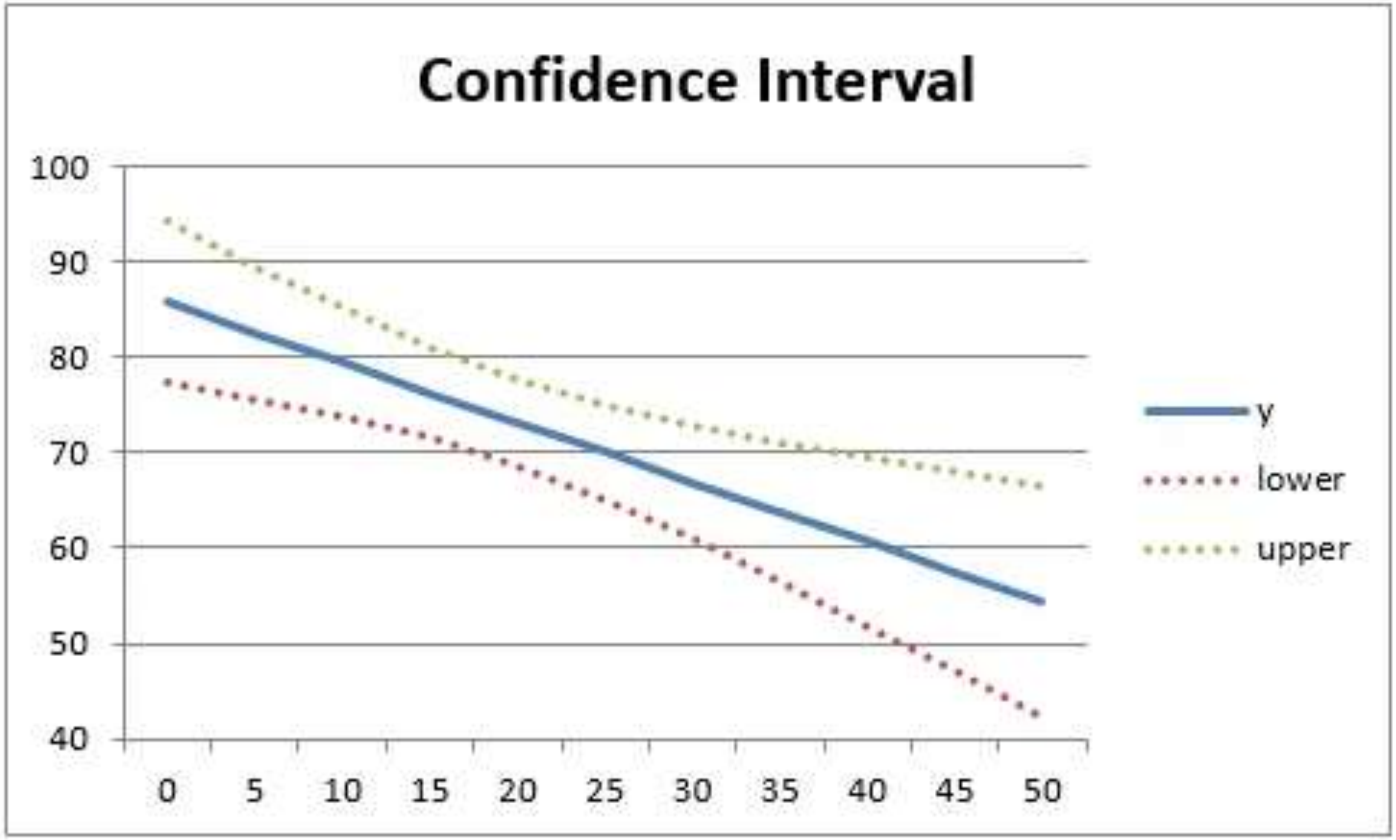
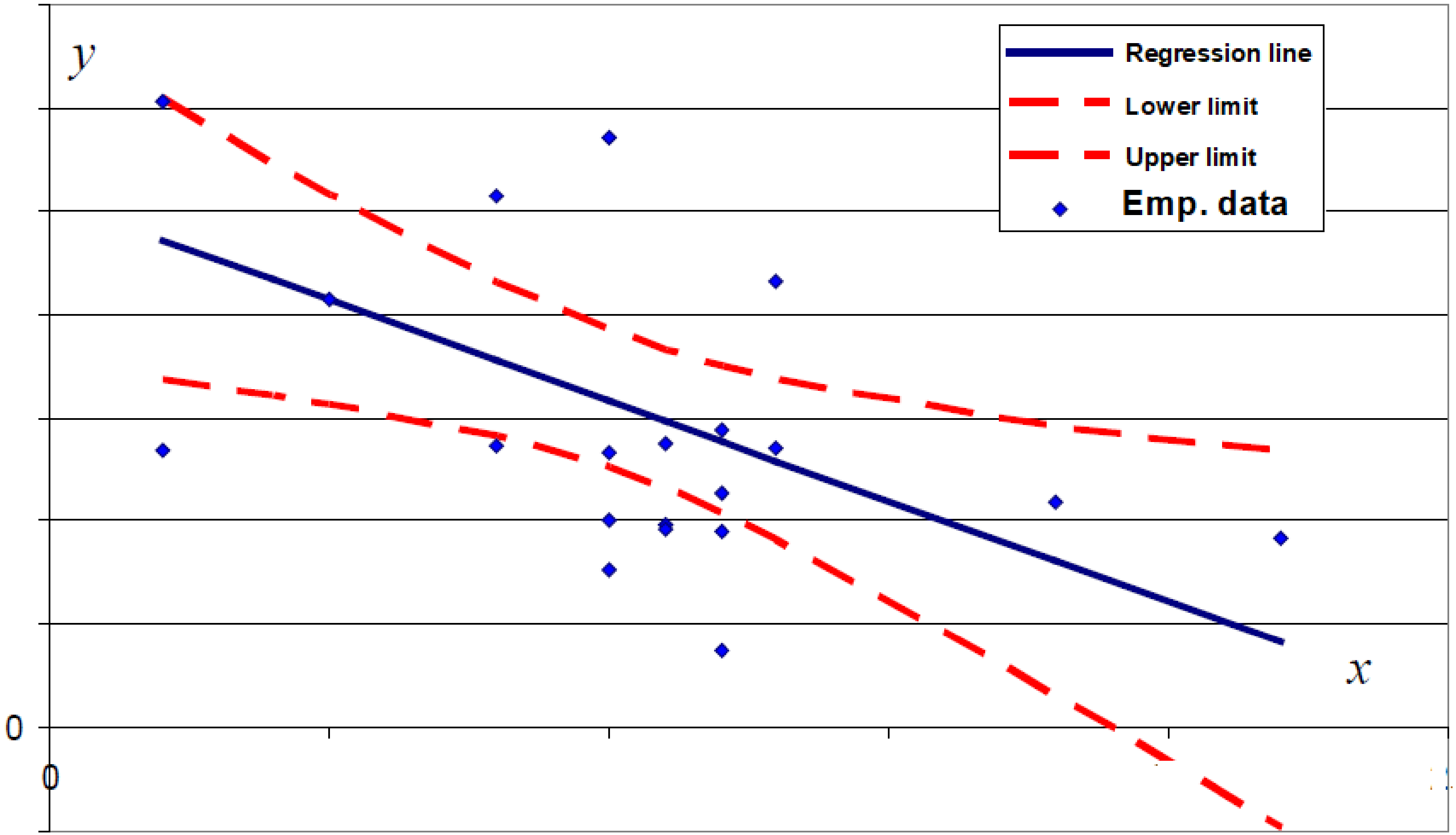


Figure – Regression confidence interval chart

Confidence interval with the significance level α



Confidence interval with the significance level α

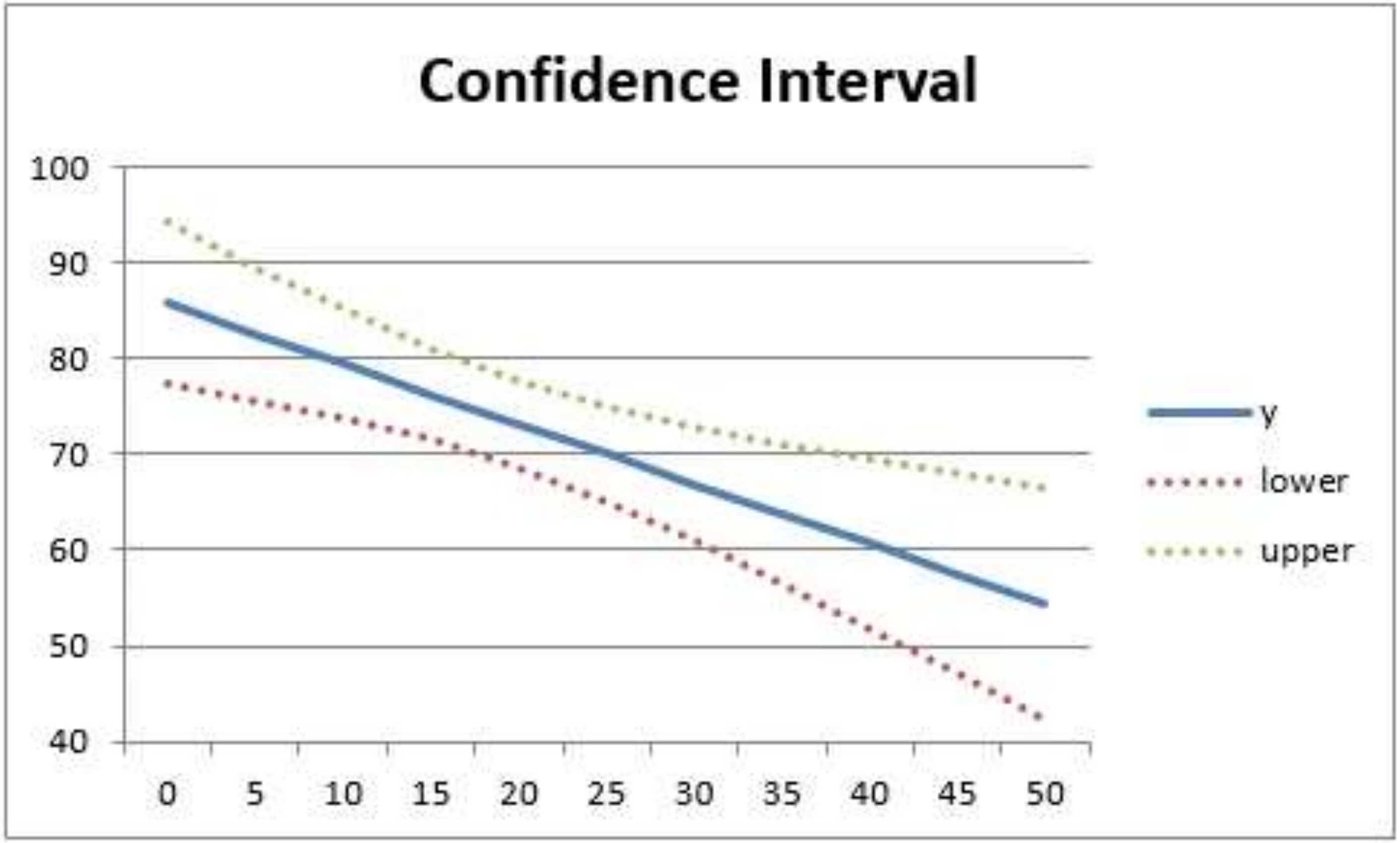


Figure – Regression confidence interval chart

INTERVAL ESTIMATIONS

Here γ is the **confidence probability (or level)** ;

$\alpha = 1 - \gamma$ is the **significance level**.

Definition

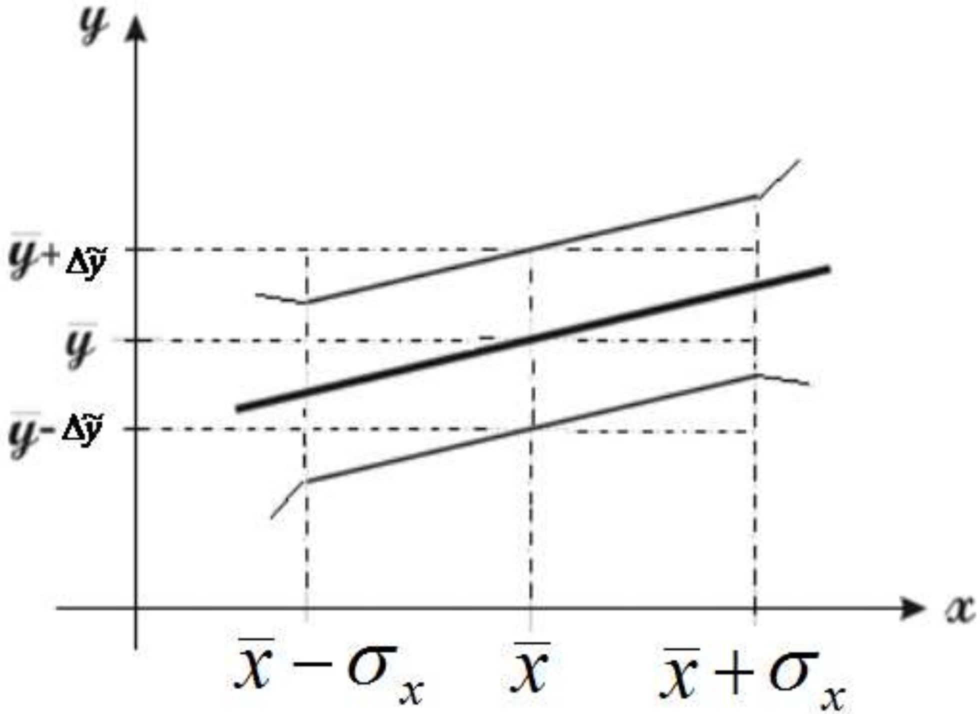
Confidence Level : 1 – α

The **probability** that the confidence interval actually contains **empirical data**.

The most common confidence levels used are **95%**, **99%**

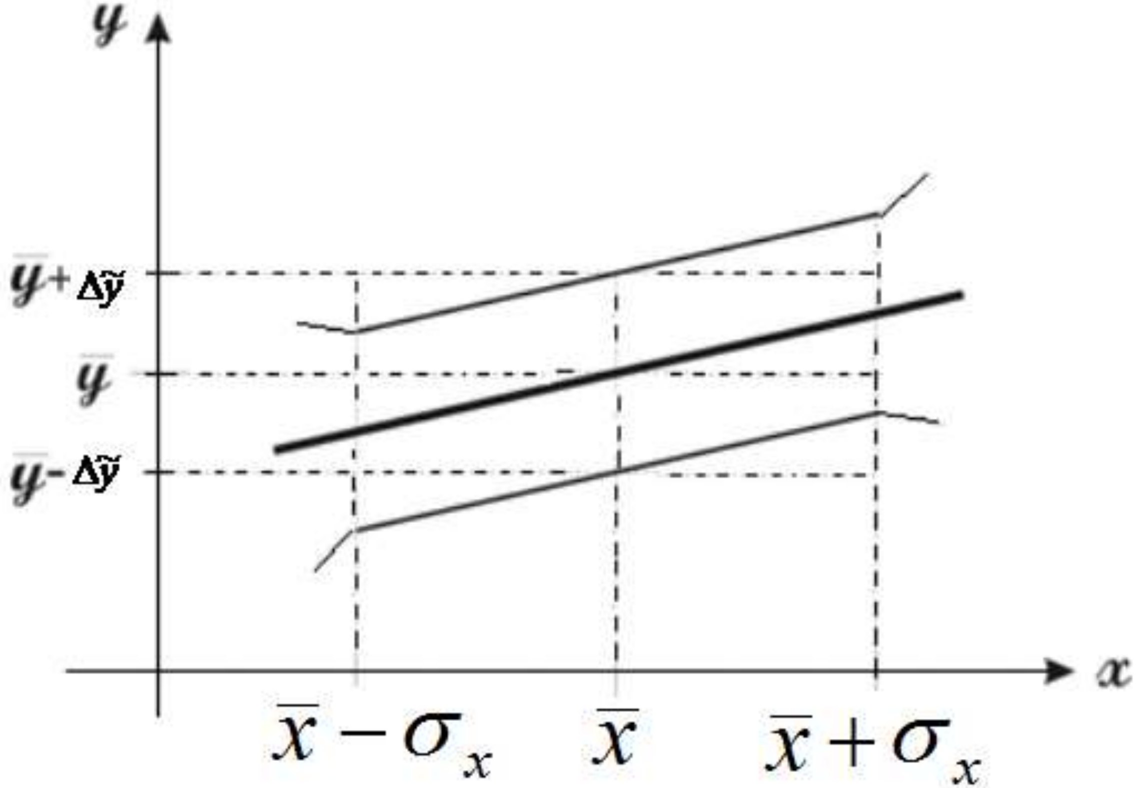
95% : **α**=0.05   99% : **α**=0.01

Confidence interval with α



$$\Delta\tilde{y}(x) = t_\alpha(n-2)\cdot\frac{\sigma_{res}}{\sqrt{n}}\cdot\sqrt{1+\frac{\left(x-\bar{x}\right)^2}{\sigma_x^2}}$$

Confidence interval with 95%



$$\Delta\tilde{y} \approx t_{0,05}(n-2)\cdot\sigma_y\cdot\sqrt{\frac{1-r^2}{n-2}}$$

Confidence interval with α

$$\left(\tilde{y}_i - \Delta\tilde{y}_i ; \tilde{y}_i + \Delta\tilde{y}_i\right)$$

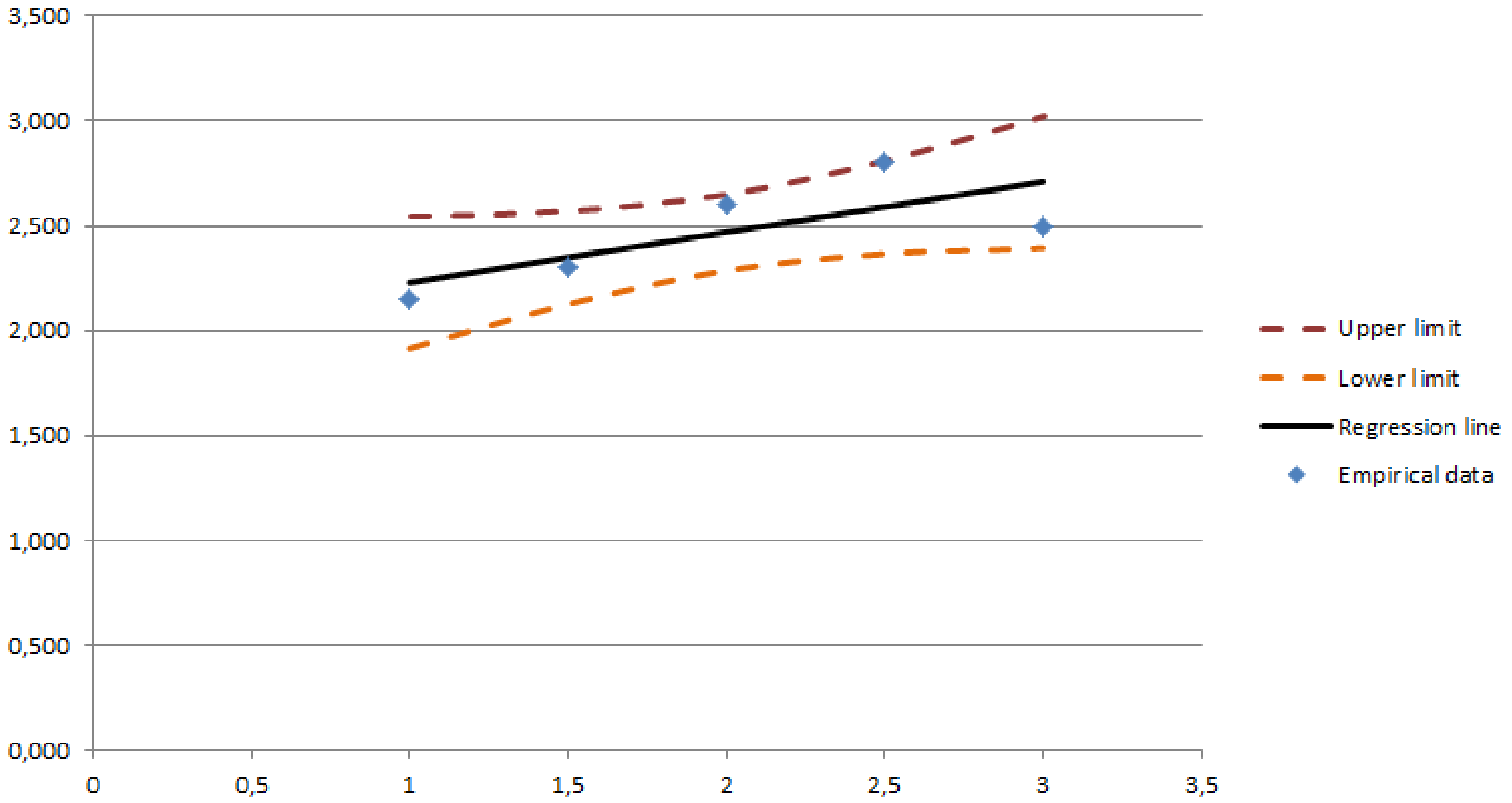
$$\Delta\tilde{y}(x) = t_\alpha(n-2) \cdot \frac{\sigma_{res}}{\sqrt{n}} \cdot \sqrt{1 + \frac{\left(x - \bar{x}\right)^2}{\sigma_x^2}}$$

Confidence interval with $\alpha = 0,05$

	x	y	y th	(xi -x mean)^2	x^2	Δy th	y th - Δ y th	y th + Δ y th
1	1	2,15	2,23	1,00	1	0,31604	1,914	2,546
2	1,5	2,3	2,35	0,25	2,25	0,223474	2,127	2,573
3	2	2,6	2,47	0,00	4	0,182466	2,288	2,652
4	2,5	2,8	2,59	0,25	6,25	0,223474	2,367	2,813
5	3	2,5	2,71	1,00	9	0,31604	2,394	3,026
sum	10	12,35			22,5			
mean	2	2,47			4,5			

b1	0,24	1,99	b0
m b1	0,123288	0,261534	m b0
R ^2	0,55814	0,194936	Sigra res
F emp	3,789474	3	n-2
SSR	0,144	0,114	SSE
t 0,05 (3)	3,18		
sigma x ^2	0,5		

Confidence interval with $\alpha = 0,05$



ELASTICITY

The elasticity coefficient is a number that indicates the percentage change that will occur in one variable (*y*) when the variable *x* changes one percent.

$$\overline{\mathrm{E}} = f'(x) \cdot \frac{\overline{x}}{\overline{\overline{y}}} = b_1 \cdot \frac{\overline{x}}{\overline{\overline{y}}}$$

ELASTICITY

The elasticity coefficient indicates the influence of X on Y.

$$\overline{\mathrm{E}} = b_1 \cdot \frac{\overline{x}}{\overline{y}} = 0{,}24 \cdot \frac{2}{2{,}47} = 0{,}2\%$$

Thank You for your time and attention! Hope this presentation was educational and helpful to you ☺