GDAŃSK UNIVERSITY
OF TECHNOLOGY

FACULTY OF MANAGEMENT AND ECONOMICS

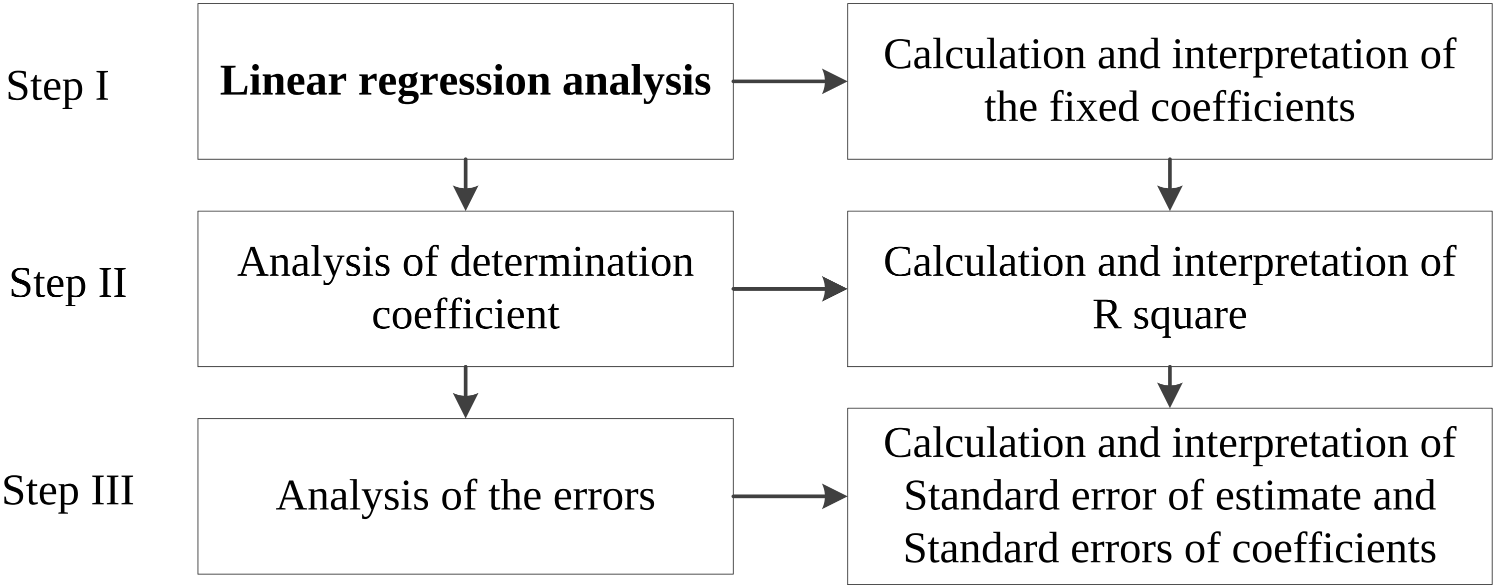
LINEAR REGRESSION

KAROLINA TURA, PHD
DEPARTMENT OF ECONOMIC SCIENCE

AGENDA

1. **Linear regression**
2. **R square**
3. **Analysis of the errors**
4. **Practice**

STEPS OF THE ANALYSIS



Step I

Linear regression analysis

Calculation and interpretation of the fixed coefficients

Step II

Analysis of determination coefficient

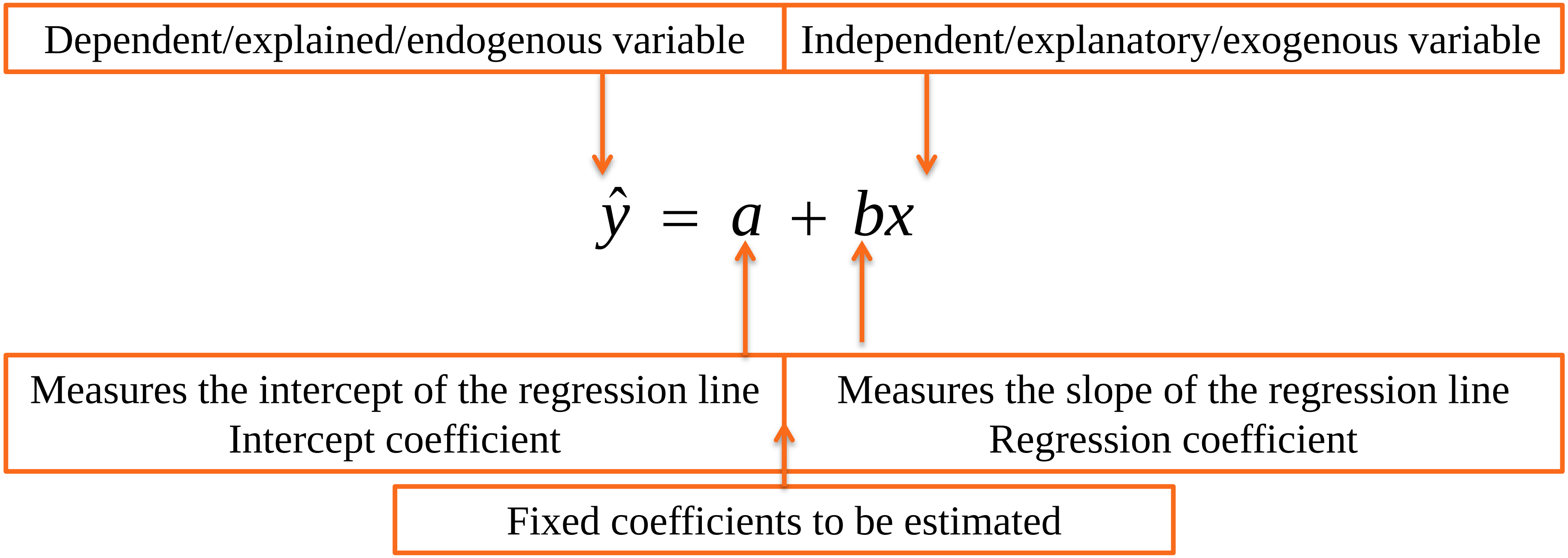
Calculation and interpretation of R square

Step III

Analysis of the errors

Calculation and interpretation of Standard error of estimate and Standard errors of coefficients

1. REGRESSION LINE



Dependent/explained/endogenous variable

Independent/explanatory/exogenous variable

$$\hat{y} = a + bx$$

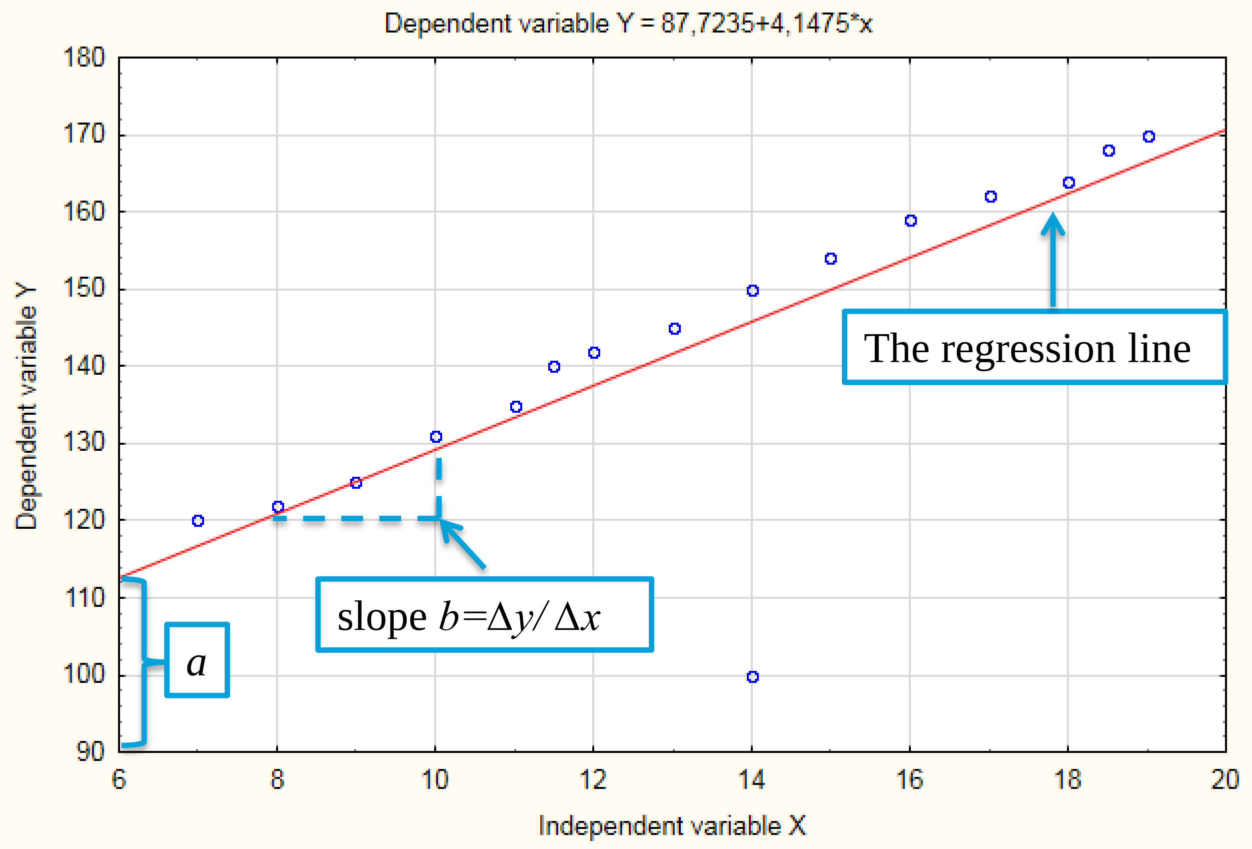
Measures the intercept of the regression line
Intercept coefficient

Measures the slope of the regression line
Regression coefficient

Fixed coefficients to be estimated

The way of examining the relationship between two or more variables

1. LINEAR REGRESSION – GEOMETRIC INTERPRETATION



$$\hat{y} = a + bx$$

Regression analysis describes this causal relationship by fitting a straight line drawn through the data, which best summarises them „The line of best fit”

1. REGRESSION LINE- FORMULAS

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a = \overline{y} - b\overline{x}$$

TASK 1. REGRESSION LINE

Task 1. The table shows the birth rate and GNP growth in 12 countries. Create a scatter plot. Find and interpret linear regression line (Birth rate- dependent variable).

Country	Birth rate	GNP growth [%]
Brazil	30	5,1
Colombia	29	3,2
Costa Rica	30	3
India	35	1,4
Mexico	36	3,8
Peru	36	1
Philippines	34	2,8
Senegal	48	-0,3
South Korea	24	6,9
Sri Lanka	27	2,5
Taiwan	21	6,2
Thailand	30	4,6

HINT

Country	y_i	x_i	x_i^2	$x_i y_i$
Brazil	30	5,1	26,01	153
Colombia	29	3,2	10,24	92,8
Costa Rica	30	3	9	90
India	35	1,4	1,96	49
Mexico	36	3,8	14,44	136,8
Peru	36	1	1	36
Philippines	34	2,8	7,84	95,2
Senegal	48	-0,3	0,09	-14,4
South Korea	24	6,9	47,61	165,6
Sri Lanka	27	2,5	6,25	67,5
Taiwan	21	6,2	38,44	130,2
Thailand	30	4,6	21,16	138
Sum	380	40,2	184,04	1139,7

$$\bar{y} = \frac{380}{12} = 31.67$$

$$\bar{x} = \frac{40.2}{12} = 3.35$$

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} =$$

$$= \frac{12 * 1139.7 - 40.2 * 380}{12 * 184.04 - (40.02)^2} = -2.7$$

$$a = \bar{y} - b\bar{x} =$$

$$31.67 - 3.35 * (-2.7) = 40.71$$

The slope coefficient of b=-2.7 implies that a unit increase in the growth rate would decrease the birth rate by 2.7%.

2. COEFFICIENT OF DETERMINATION

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$$

Total sum of squares

$$ESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i$$

Error sum of squares

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = TSS - ESS$$

Regression sum of squares

$$R^2 = \frac{RSS}{TSS}$$

$$R^2 \in < 0,1 >$$

R square- the measure of goodness of fit

R^2	Interpretation
1	Indicates that all sample observations lie exactly on the regression line
0	Indicates that the regression line is of no use at all

TASK 2. R SQUARE

Task 2. The table shows the birth rate and GNP growth in 12 countries. Find and interpret linear regression line (Birth rate- dependent variable) and determination coefficient.

Country	Birth rate	GNP growth [%]
Brazil	30	5,1
Colombia	29	3,2
Costa Rica	30	3
India	35	1,4
Mexico	36	3,8
Peru	36	1
Philippines	34	2,8
Senegal	48	-0,3
South Korea	24	6,9
Sri Lanka	27	2,5
Taiwan	21	6,2
Thailand	30	4,6

HINT

Country	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
Brazil	30	5,1	26,01	153	900
Colombia	29	3,2	10,24	92,8	841
Costa Rica	30	3	9	90	900
India	35	1,4	1,96	49	1225
Mexico	36	3,8	14,44	136,8	1296
Peru	36	1	1	36	1296
Philippines	34	2,8	7,84	95,2	1156
Senegal	48	-0,3	0,09	-14,4	2304
South Korea	24	6,9	47,61	165,6	576
Sri Lanka	27	2,5	6,25	67,5	729
Taiwan	21	6,2	38,44	130,2	441
Thailand	30	4,6	21,16	138	900
Sum	380	40,2	184,04	1139,7	12564

$$\bar{y} = \frac{380}{12} = 31.67 \qquad \bar{x} = \frac{40.2}{12} = 3.35$$

$$TSS = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 =$$

$$= 12564 - 12 * (31.67)^2 = 530.667$$

$$ESS = \sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i =$$

$$= 12564 - 40.71 * 380 - (-2.7) * 1139.7 =$$

$$= 170.75$$

$$RSS = TSS - ESS =$$

$$= 530.67 - 170.75 = 359.91$$

$$R^2 = \frac{RSS}{TSS} = \frac{359.91}{530.67} \approx 0.67$$

67.8% of variation of Birth rate around the overall mean is expalined by the variation in countries growth rates

3. ANALYSIS OF THE ERRORS

Standard error of estimate

$$\hat{y} = \underset{(S(a))}{a} + \underset{(S(b))}{b}\; x +/- S_{y}$$

Standard error of the coefficient *a*

Standard error of the coefficient *b*

3. ANALYSIS OF THE ERRORS- FORMULAS

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-k}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-k}}$$

Standard error of estimate

$$S(b) = \frac{S_y}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}}$$

Standard error of the coefficient *b*

$$S(a) = \sqrt{\frac{S_y^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)}}$$

Standard error of the coefficient *a*

TASK 3. ANALYSIS OF ERRORS

Task 3. The table shows the birth rate and GNP growth in 12 countries. Find and interpret linear regression line (Birth rate- dependent variable) and errors.

Country	Birth rate	GNP growth [%]
Brazil	30	5,1
Colombia	29	3,2
Costa Rica	30	3
India	35	1,4
Mexico	36	3,8
Peru	36	1
Philippines	34	2,8
Senegal	48	-0,3
South Korea	24	6,9
Sri Lanka	27	2,5
Taiwan	21	6,2
Thailand	30	4,6

HINT

Country	y_i	x_i	x_i^2	$x_i y_i$	y_i^2
Brazil	30	5,1	26,01	153	900
Colombia	29	3,2	10,24	92,8	841
Costa Rica	30	3	9	90	900
India	35	1,4	1,96	49	1225
Mexico	36	3,8	14,44	136,8	1296
Peru	36	1	1	36	1296
Philippines	34	2,8	7,84	95,2	1156
Senegal	48	-0,3	0,09	-14,4	2304
South Korea	24	6,9	47,61	165,6	576
Sri Lanka	27	2,5	6,25	67,5	729
Taiwan	21	6,2	38,44	130,2	441
Thailand	30	4,6	21,16	138	900
Sum	380	40,2	184,04	1139,7	12564

$$S_Y = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - a\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i y_i}{n-k}} =$$

$$= \sqrt{\frac{12564 - 40.71 * 380 - (-2.7) * 1139.7}{12-2}} = 4.13$$

$$S(b) = \frac{S_y}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}} =$$

$$= \frac{4.13}{\sqrt{184.04 - 12 * 3.35^2}} = 0.59$$

$$S(a) = \sqrt{\frac{S_y^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}} =$$

$$= \sqrt{\frac{4.13^2 * 184.04}{12 * (184.04 - 12 * 3.35^2)}} = 2.3$$

Theoretical values of Birth rate differ from the empirical ones by +/- 4.13 on average.
Estimating the intercept coefficient we are making mistakes by +/- 2.3 on average.
Estimating the slope coefficient we are making mistakes by +/- 0.59 on average.

STATISTICA

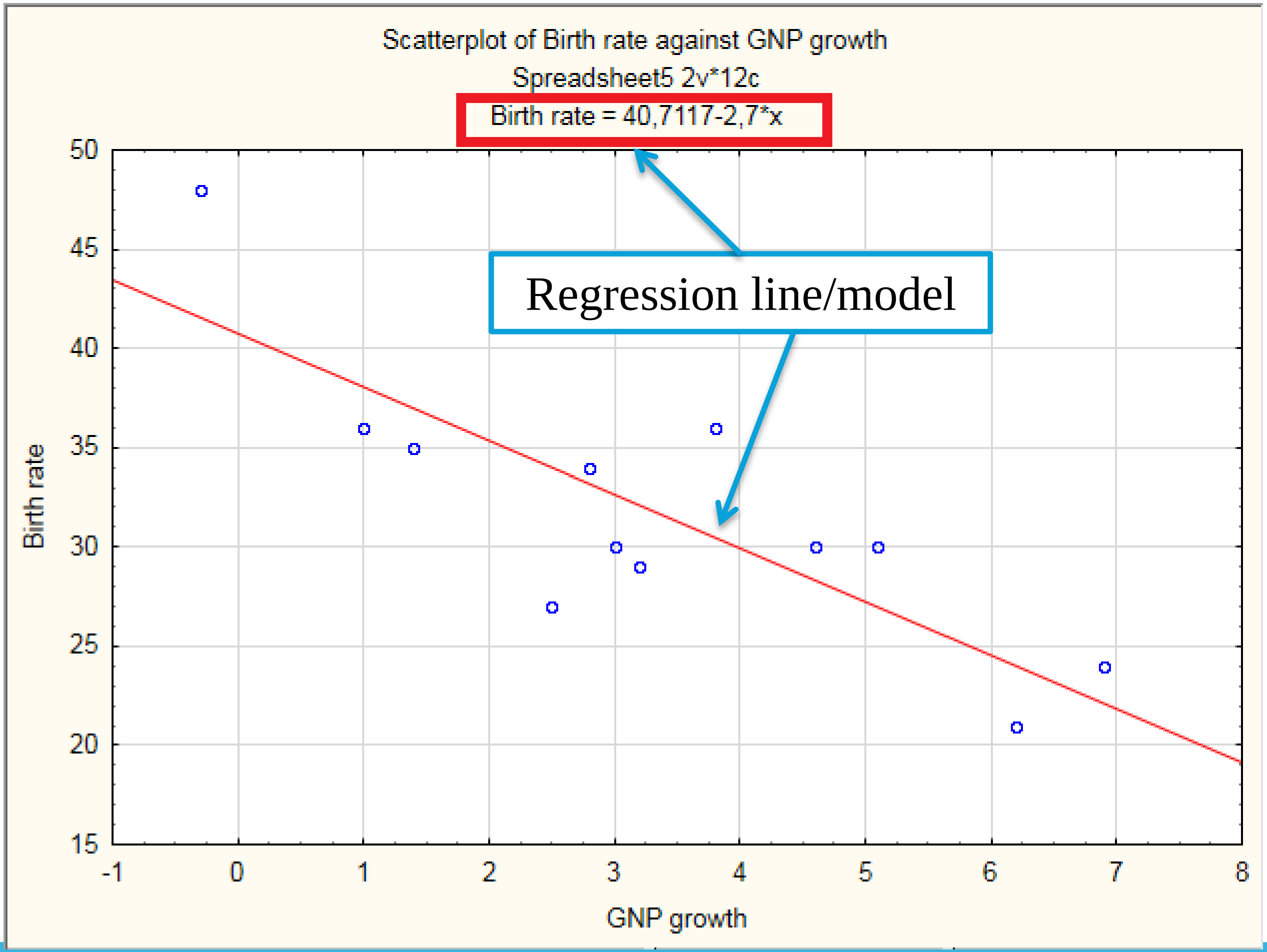
TASK 4.

Task 4. The table shows the birth rate and GNP growth in 12 countries. Create a scatter plot. Find and interpret linear regression line (Birth rate- dependent variable), R square and errors. Predict the birth rate when GNP growth is 2 %.

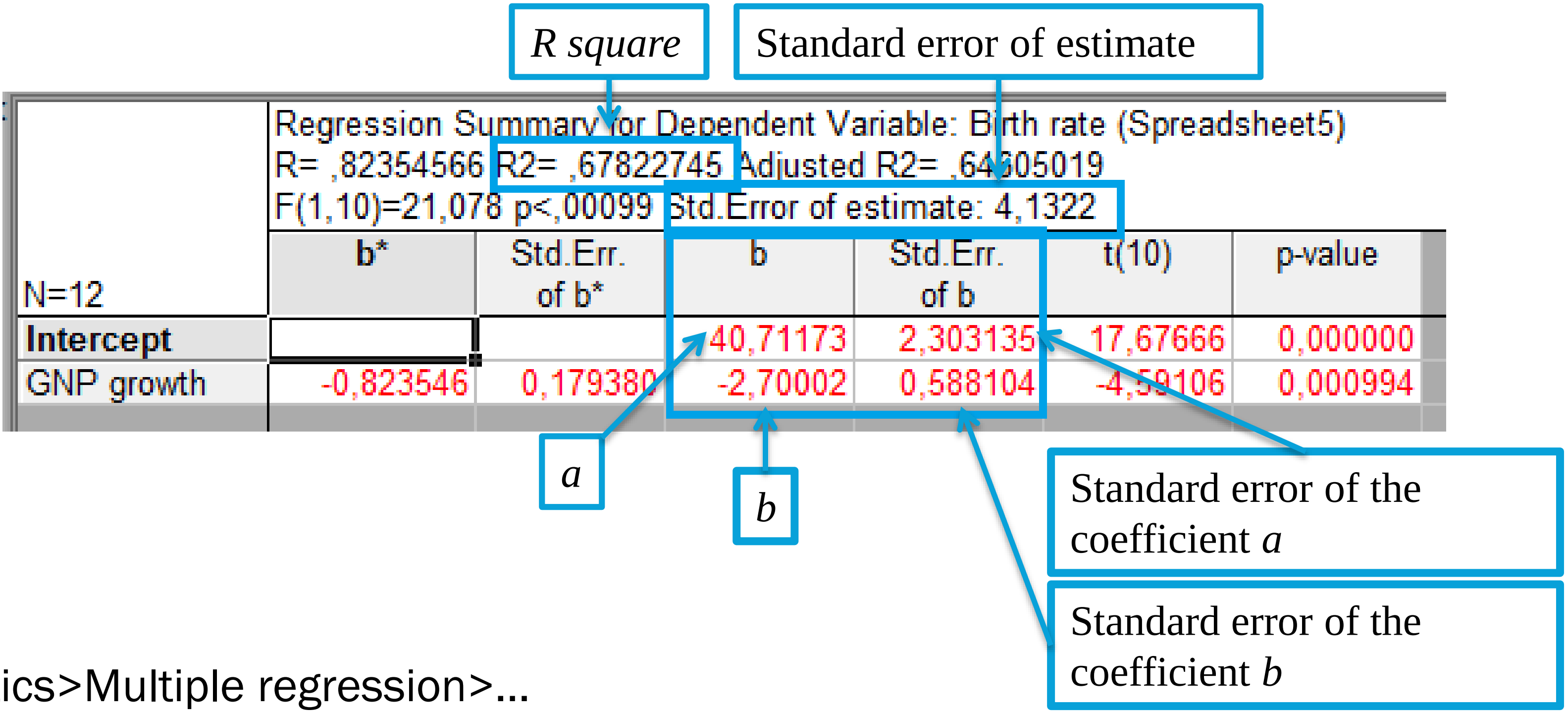
Country	Birth rate	GNP growth [%]
Brazil	30	5,1
Colombia	29	3,2
Costa Rica	30	3
India	35	1,4
Mexico	36	3,8
Peru	36	1
Philippines	34	2,8
Senegal	48	-0,3
South Korea	24	6,9
Sri Lanka	27	2,5
Taiwan	21	6,2
Thailand	30	4,6

HINT

Graphs>Scatterplots>



HINT


R square
Standard error of estimate
a
b
Standard error of the coefficient *a*
Standard error of the coefficient *b*

Regression Summary for Dependent Variable: Birth rate (Spreadsheet5)
R= ,82354566 R2= ,67822745 Adjusted R2= ,64605019
F(1,10)=21,078 p<,00099 Std.Error of estimate: 4,1322

N=12	b*	Std.Err. of b*	b	Std.Err. of b	t(10)	p-value
Intercept			40,71173	2,303135	17,67666	0,000000
GNP growth	-0,823546	0,179380	-2,70002	0,588104	-4,59106	0,000994

Statistics>Multiple regression>...

MULTIPLE LINEAR REGRESSION

$$\hat{y} = \underset{(S(a))}{a} + \underset{(S(b_1))}{b_1} x_1 + \underset{(S(b_2))}{b_2} x_2 + \ldots + \underset{(S(b_n))}{b_n} x_n + / - S_y$$

TASK.5.

A real estate agent would like to predict the selling price of single-family homes. After careful consideration, he concludes that the variables likely to be most closely related to selling price are: the size of the house (in 100s ft^2). and the age of the house. As an experiment, he takes a random sample of fifteen recently sold houses and records the selling price (in $ 1,000s). These are shown in the accompanying table. Find and interpret the linear regression model (Dependent variable- Selling Price). Predict the selling price when: house size is 100, age- 10.

TASK 5.

House size	Selling Price	Age (years)
20	89,5	5
14,8	79,9	10
20,5	83,1	8
12,5	56,9	7
18	66,6	8
14,3	82,5	12
27,5	126,3	1
16,5	79,3	10
24,3	119,9	2
20,2	87,6	8
22	112,6	7
19	120,8	11
12,3	78,5	16
14	74,3	12
16,7	74,8	13

TASK 5.

$$\hat{y} = \underset{(34.06)}{-25.58} + \underset{(1.3)}{5.35}\, x_1 + \underset{(1.41)}{1.98}\, x_2 + / - 12.55$$

$$R^2 = 0.7$$

In this model, for each additional 100 square feet, the price of house increases on average by 5.35$ (assuming that the other independent variables are fixed).

In this model, for each additional year in the age of the house, the price increases on average by 1.98 (assuming that the other independent variables are fixed).

TASK 5.

$$\hat{y} = \underset{(34.06)}{-25.58} + \underset{(1.3)}{5.35}\,x_1 + \underset{(1.41)}{1.98}\,x_2 + / - 12.55$$

$$R^2 = 0.7$$

Theoretical values of Selling Price differ from the empirical ones by +/- 12.55 on average.
Estimating the intercept coefficient we are making mistakes by +/- 34.06 on average.
Estimating the coefficient b1 we are making mistakes by +/- 1.3 on average.
Estimating the coefficient b2 we are making mistakes by +/- 1.41 on average.
70% of selling prices were explained by the model.

PREPARATION FOR THE NEXT CLASSES

McClave, J. T., Benson, P. G., Sincich, T. (2008) , *Statistics for Business & Economics*, Pearson Education Inc., New Jersey.

Thank you for your attention

GDAŃSK UNIVERSITY OF TECHNOLOGY

FACULTY OF MANAGEMENT AND ECONOMICS