

## Biometry practical 6

### Illustrated (imperfect) practical guide

#### Preparatory work

1. Open in *MS Excel* the questionnaire data (file analysed already in previous practical),
  2. insert new worksheet, rename it to 'Praks6' (or 'Practical6'),
  3. and make a copy of the data table (from worksheet 'Andmed'/'Data') and paste it into the upper left corner of the new worksheet.
- 

The task of today's practical is to predict students' height by their shoe size.

#### Exercise 1.

- Illustrate the relationship between variables 'HEIGHT' and 'SHOE\_SIZE' with scatterplot.
- Does the x-axis (horizontal axis) corresponds to shoe size and y-axis (vertical axis) to height? If not, switch the axes.
- To predict the height by shoe size, add into scatterplot linear regression line, corresponding regression equation and determination coefficient  $R^2$  (which describes the prediction accuracy).
- Calculate based on estimated equation, what is the expected height of student with shoe size 40.

#### Exercise 2.

- Perform linear regression analysis to predict students' height by their shoe size with statistical procedure *Regression (Data-tab -> Data analysis...)*.
  - Write down the regression equation (regression model) in the form
$$\text{Height} = a + b \times \text{Shoe.size},$$
using instead of letters  $a$  and  $b$  their estimated numerical values.
  - How big is the expected heights' difference of students with shoe sizes differing by 2 units?
  - Is the estimated regression equation statistically significant? Why (based on which numbers) you made this decision?
  - Formulate one sentence about prediction accuracy using multiple correlation coefficient ( $R$ ), determination coefficient ( $R^2$ ) or model standard error (or all these characteristics).
-

## Exercise 1 guide

1. Illustrate the relationship between variables 'HEIGHT' and 'SHOE\_SIZE' with scatterplot.

**The shoe size must be on x-axis and the height on y-axis. If it is not so, switch the axes (look at the following scheme).**

By default Excel puts the first variable (HEIGHT) into x-axis. But to predict the height, it must be on y-axis (y is the dependent variable). To switch the axis ...

**Select Data Source**

Chart data range: `= 'Praks 6'!$B$1:$B$55, 'Praks 6'!$E$1:$E$55`

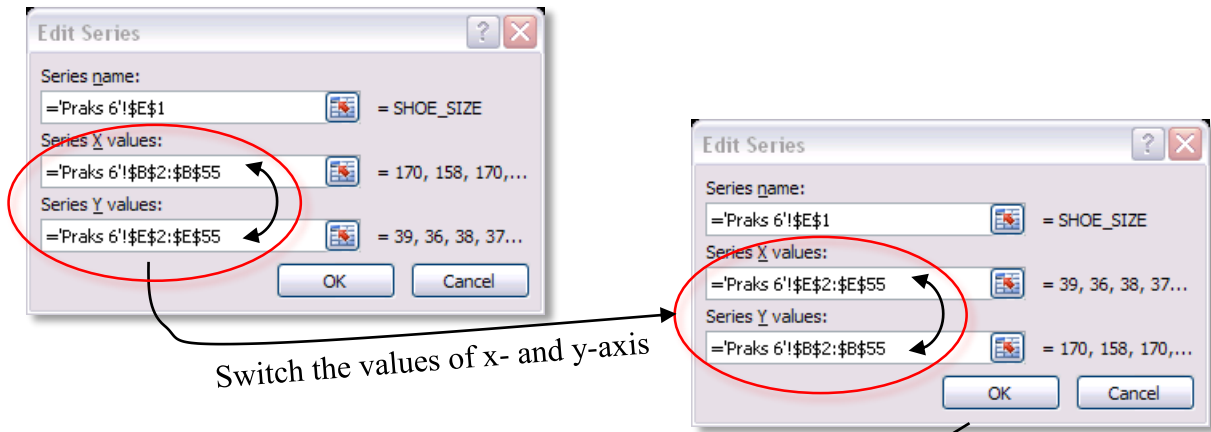
Switch Row/Column

Legend Entries (Series)

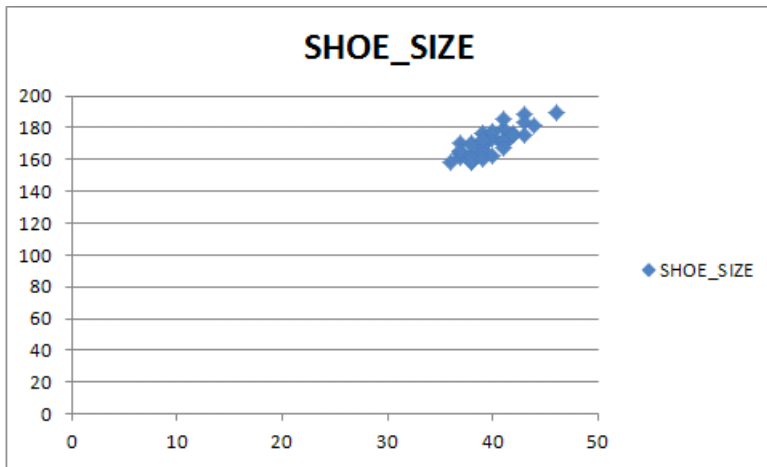
- SHOE\_SIZE

Horizontal (Category) Axis Labels

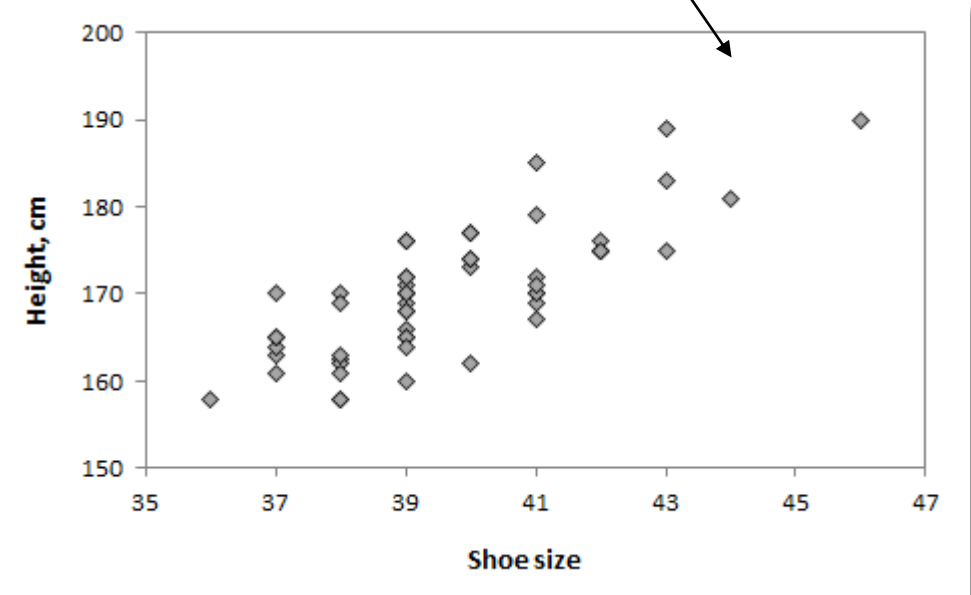
- 170
- 158
- 170
- 170
- 179



Result:  
the y-axis correspond to height  
and the x-axis to shoe size.



Format the diagram!



2. To predict the height by shoe size, add into scatterplot **linear regression line**.

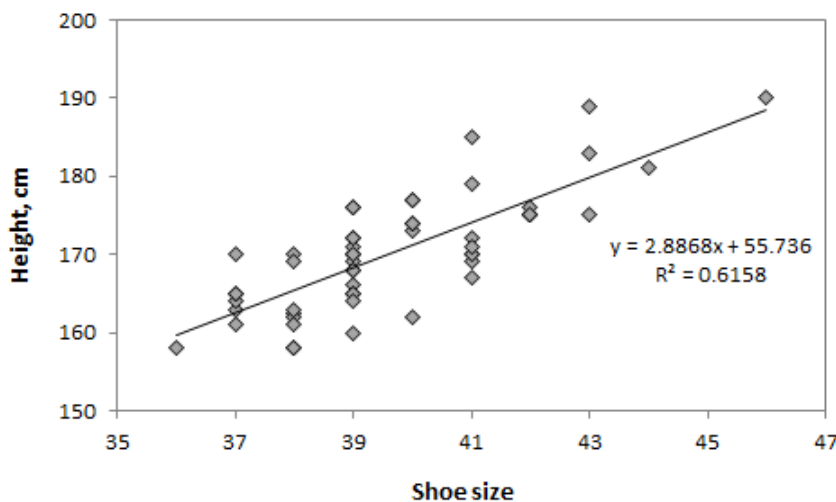
Add into diagram also **regression equation** and **determination coefficient  $R^2$**  (which describes the prediction accuracy).

Two variants to add linear trendline

18	0	no
19	0	yes
20	2.5	yes
21	1	yes

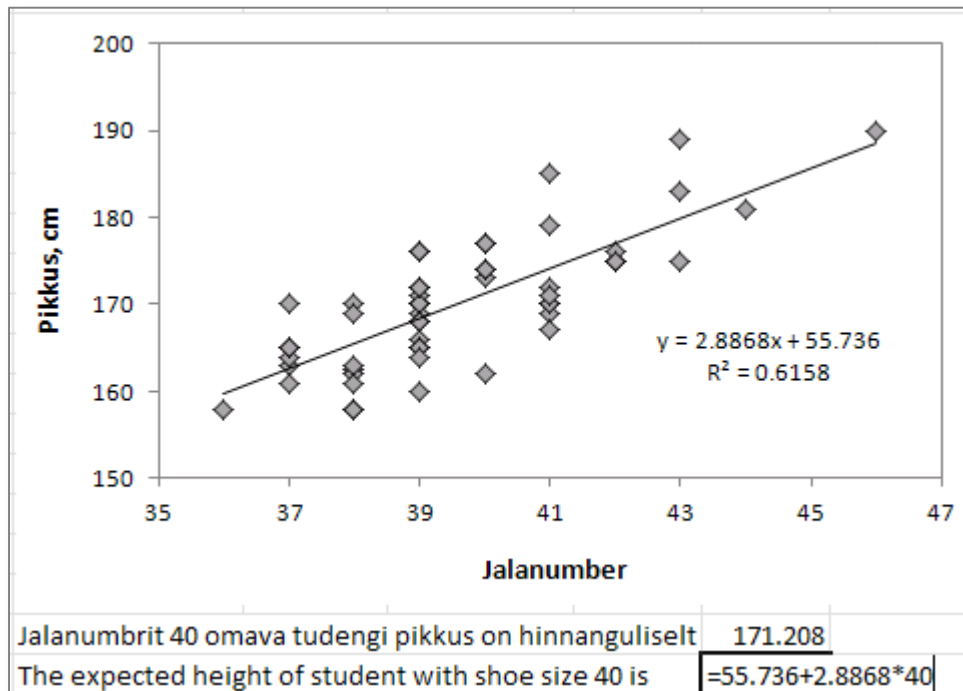
To add equation and  $R^2$

Result:



3. Predict based on estimated equation, how tall is expectably the student with shoe size 40.

To do this, type the regression equation into Excel worksheet cell (started with equation sign =) and replace argument  $x$  with value 40.



## Exercise 2 guide.

1. Perform the linear regression analysis to predict students' height by their shoe size with statistical procedure *Regression (Data-tab -> Data analysis...)*.

The screenshot shows the following data in the spreadsheet:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
1	GENDER	HEIGHT	WEIGHT	HEAD	SHOE_S	MATH	SMOKE																		
2	W	170	70	55.5	39	3	no																		
3	W	158	47.5	55	36	3	no																		
4	W	170	60	53	38	5	no																		
5	W	170	50	55	37	4	no																		
6	W	179	68	58	41	5	no																		
7	W	163	56		37	4	no																		
8	W	177	65	55	40	3	no																		
9	W	162.5	53	55	38	3	no																		
10	W	170	75	56	39	5	no																		
11	M	175	74	57	42	3	no																		
12	W	176	66	57	39	4	no																		
13	M	175	64	56	42	4	no anymore, b																		
14	M	190	82	58	46	4	no																		
15	W	161	50	55	37	4	no																		
16	W	170	85	57	41	4	no																		
17	W	176	58	52	39	5	no																		
18	W	172	90	58	41	4	no																		
19	W	158	55	57	38	4	yes																		
20	M	189	82		43	4	yes																		
21	W	169	60	55.5	41	4	yes																		
22	W	164	52	56	37	4	no																		
23	W	172	62	56	39	4	no																		
24	W	173	66	56	40	5	no																		
25	W	169	60	55	39	3	no																		
26	W	162	50	50	38	3	no																		
27	W	165	52	50.5	37	3	no																		
28	M	170	80	56	41	4	no																		
29	M	176	74	56	42	4	no																		
30	M	175	73	54	43	4	no																		
31	W	171	63	57	39	3	no																		
32	W	170	60	53	39	3	no																		
33	W	163	62	55	38	3	no																		
34	M	181	74	55	44	4	no																		
35	W	168	60	55	39	3	no																		
36	W	174	54	55	40	4	no																		
37	W	166	68	56	39	3	no																		
38	W	168	63	53	39	3	no																		
39	W	165	58	56	37	3	no																		
40	W	171	75	55	41	4	no																		
41	W	165	77	58	39	3	no																		
42	W	161	55	57	38	3	no																		
43	M	183	75		43	4	no																		
44	W	169	53	55	38	3	no																		
45	W	175	60	57	42	4	no																		
46	W	167	80	57.5	41	4	no																		
47	W	158	70	55	38	3	no																		
48	M	174	87	57	40	4	no																		
49	W	165	61	57	39	3	no																		
50	W	164	58	57	39	3	no																		
51	W	185	80	60	41	4	no																		
52	W	177	63	60	40	4	no																		
53	W	160	70	57	39	3	no																		
54	W	162	70	55	40	3	no																		
55	W	172	58	62	39	4	no anymore, but I've smoked																		

Result of regression analysis:

SUMMARY OUTPUT						
<i>Regression Statistics</i>						
Multiple R	0.7847576					
R Square	0.6158445					
Adjusted R Square	0.6084569					
Standard Error	4.6287793					
Observations	54					
<i>ANOVA</i>						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12	
Residual	52	1114.131081	21.4256			
Total	53	2900.208333				
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	55.73624	12.55193924	4.440448	4.71E-05	30.54893075	80.923548
SHOE_SIZE	2.8868487	0.316184355	9.130271	2.18E-12	2.252378374	3.521319

2. Write down the regression equation (regression model) in the form

$$\text{Height} = a + b \times \text{Shoe.size},$$

where instead  $a$  and  $b$  are their estimated numerical values.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	55.73624	12.55193924	4.440448	4.71E-05	30.54893075	80.923548
SHOE_SIZE	2.8868487	0.316184355	9.130271	2.18E-12	2.252378374	3.521319

3. How big is the expected difference between heights' of students with shoe sizes differing by 2 units?

Answer:  $2 \times b$  (but numerically?). Write down the sentence with numerical answer.

**4. Is the estimated regression equation statistically significant? Why (based on which numbers) you made this decision?**

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12
Residual	52	1114.131081	21.4256		
Total	53	2900.208333			

= *p*

**Remark.** The statistical significance of regression model means that the estimated equation  $Height = a + b \times Shoe.size$  predicts students' height more accurately compared with the constant model  $Height = a$

(height of all students is the same (and equal to *a*) irrespective of their shoe size).

The hypothesis pair tested in ANOVA table is of the form:

$H_0$ : regression model is not statistically significant

$H_1$ : regression model is statistically significant

or

$H_0$ : the estimated equation is not better than constant equation

$H_1$ : the estimated equation is better than constant equation

or mathematically

$H_0$ :  $Height = a$

$H_1$ :  $Height = a + b \times hoe.size$

In practice it is reasonable to use only statistically significant regression equation.

**5. Formulate one sentence about prediction accuracy using multiple correlation coefficient (*R*), determination coefficient ( $R^2$ ) or model standard error (or all these characteristics).**

SUMMARY OUTPUT	
<i>Regression Statistics</i>	
Multiple R	0.7847576
R Square	0.6158445
Adjusted R Square	0.6084569
Standard Error	4.6287793
Observations	54

Multiple correlation coefficient *R* measures the correlation between observed and predicted values of dependent variable.  
Bigger is better (more accurate model)!

Determination coefficient  $R^2$  measures the proportion of total variation of independent variable explained by the model (how well observed values are replicated by the model),  $0 \leq R^2 \leq 1$ . Is usually presented in percentages. Bigger is better (more accurate model)!

Model standard error *SE* estimates the standard deviation of prediction errors. So, it measures the average difference between observed and predicted values (average prediction error).  
Smaller is better (more accurate model)!

According to the results, it can be concluded, that predicting the students height by their shoe size, the average prediction error is 4.6 cm. However, the correlation between observed and predicted values is strong (multiple correlation coefficient  $R = 0.78$ ) and 62% of total variation of students' height is explained by the model ( $R^2 = 0.62$ ).