Biometry practical 6

Illustrated (imperfect) practical guide

Preparatory work

1. Open in *MS Excel* the questionary 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

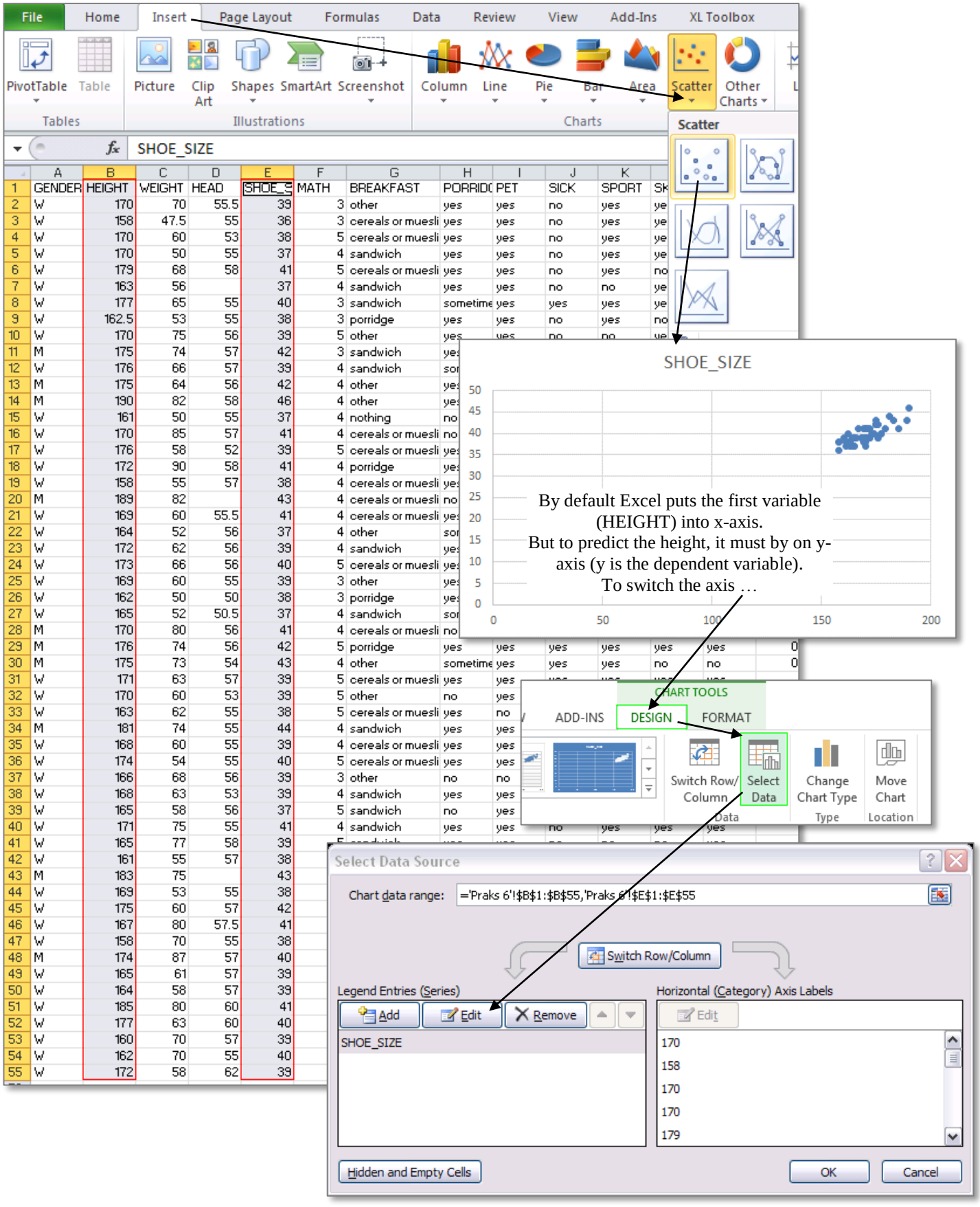
  $$\text{Height} = a + b \times \text{Shoe.size},$$

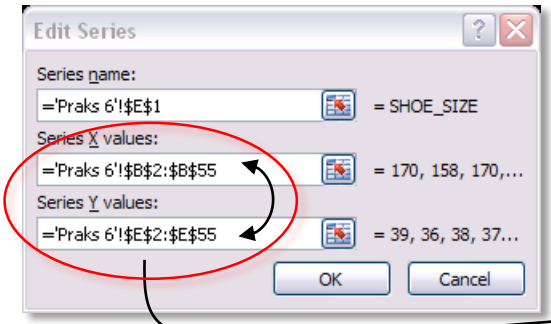
  using instead of letters a and b their estimated numerical values.
- How big is the expected heights’ difference of students with shoe sizes differing by 2 units?
- Is the estimated regression equation statistically significant? Why (based on which numbers) you made this decision?
- Formulate one sentence about prediction accuracy using multiple correlation coefficient (R), determination coefficient (R^2) or model standard error (or all these characteristics).

Exercise 1 guide

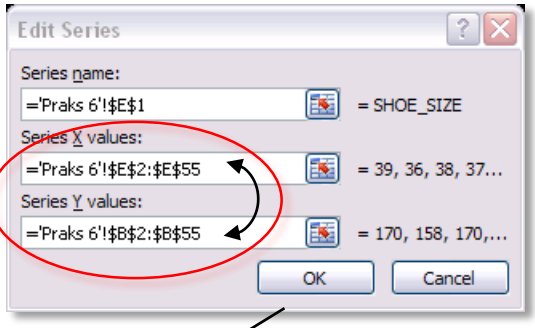
1. Illustrate the relationship between variables 'HEIGHT' and 'SHOE_SIZE' with scatterplot.

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)**.**

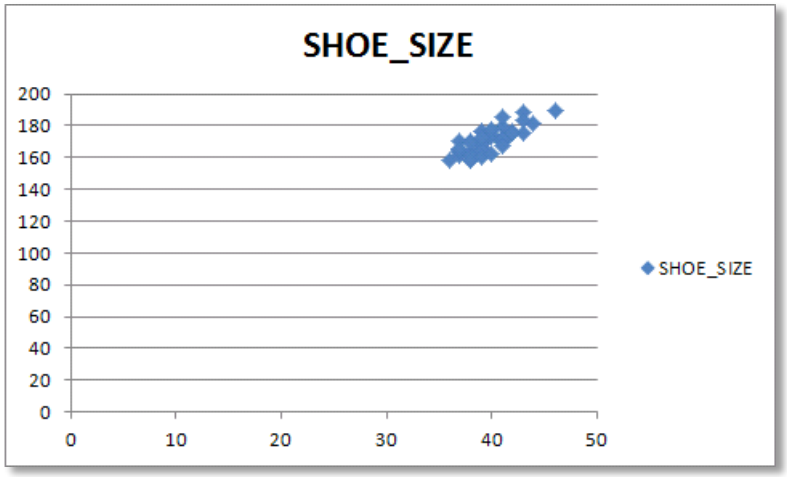




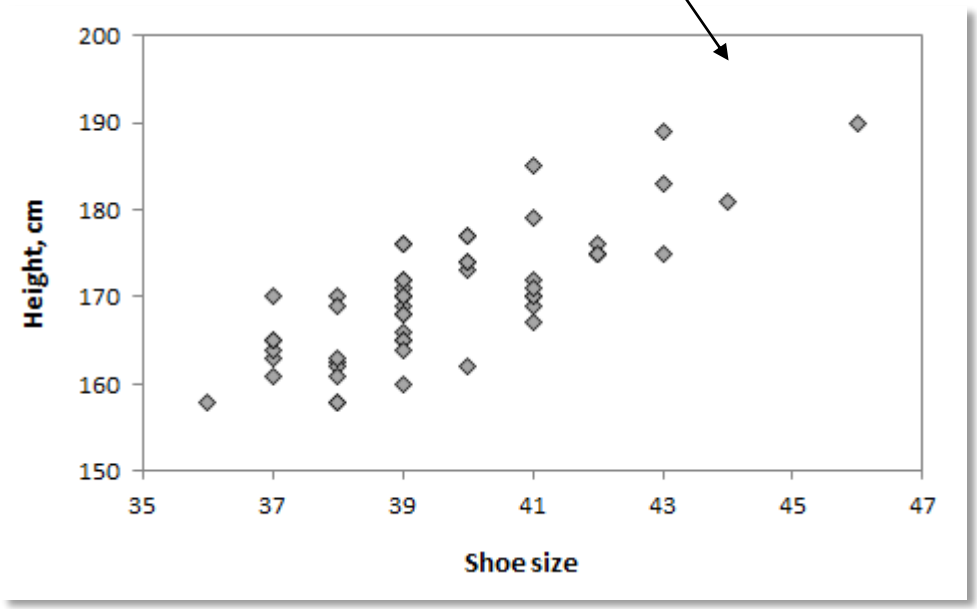
Switch the values of x- and y-axis



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

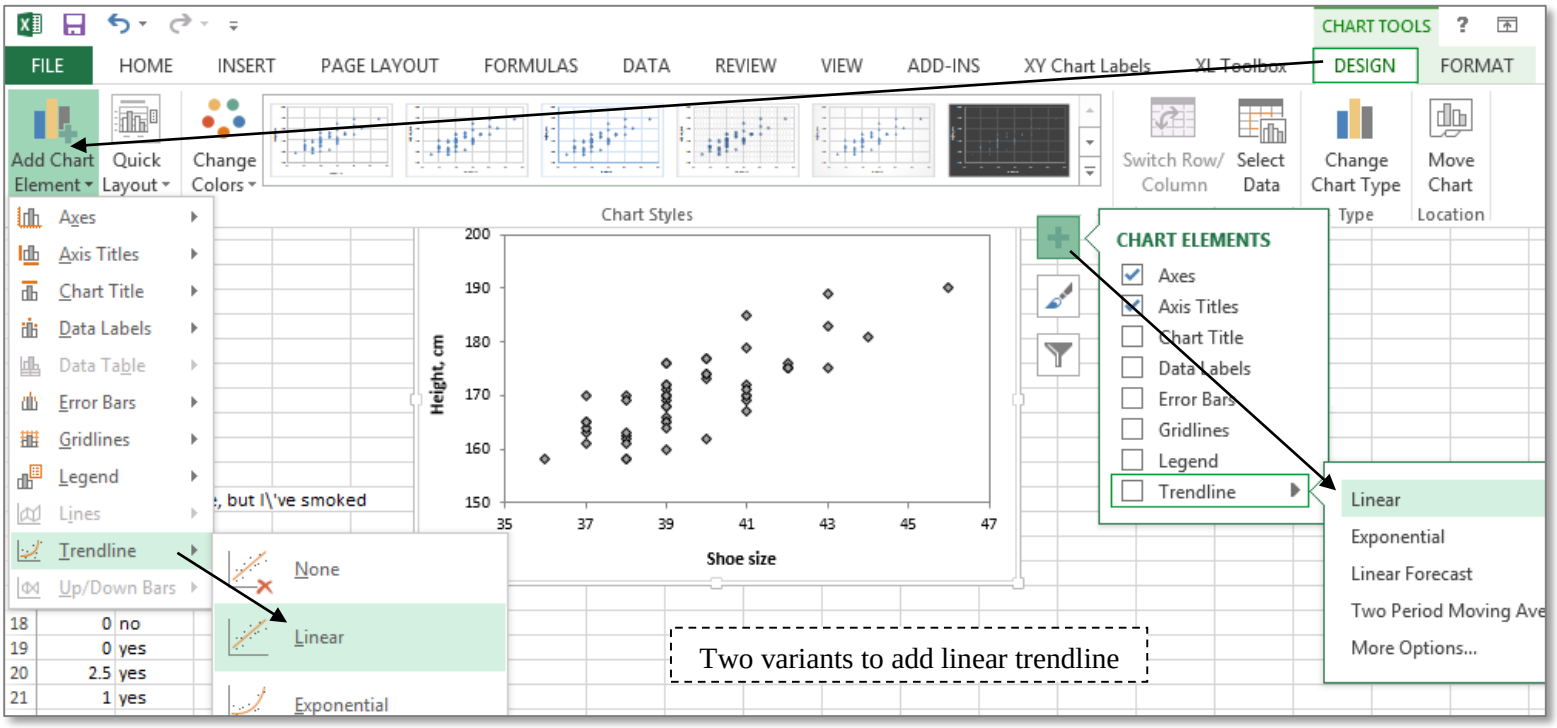


Format the diagram!

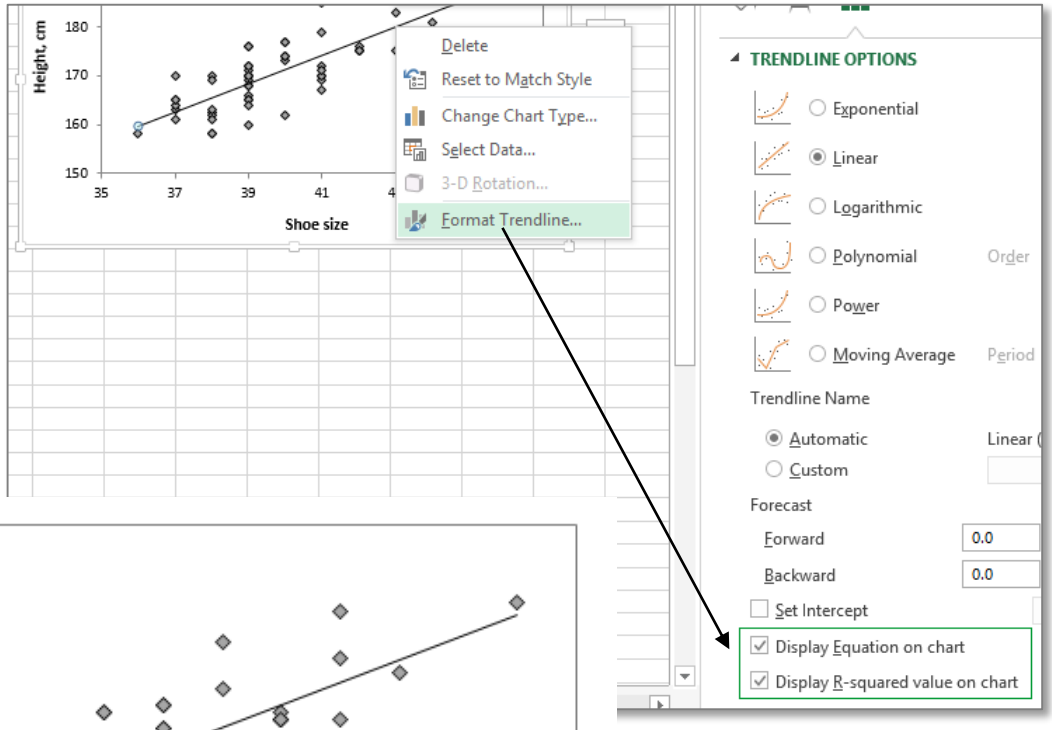


2. To predict the height by shoe size, add into scatterplot **linear regression line**.

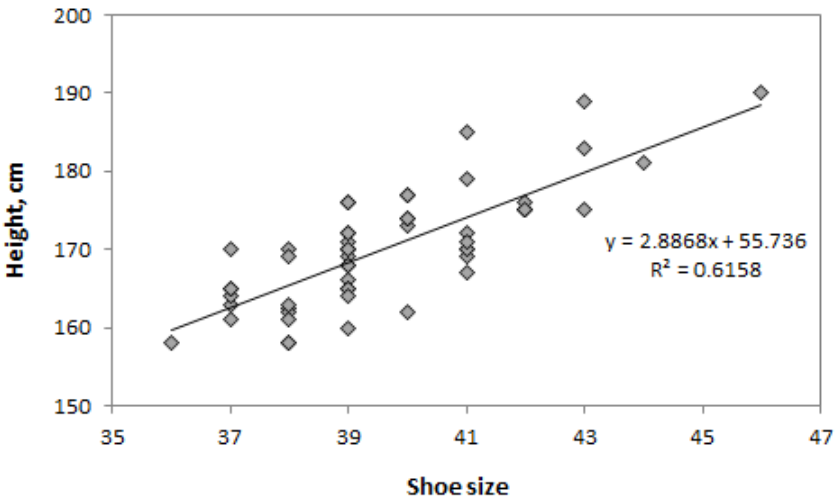
Add into diagram also **regression equation** and **determination coefficient R^2** (which describes the prediction accuracy).



To add equation and R^2

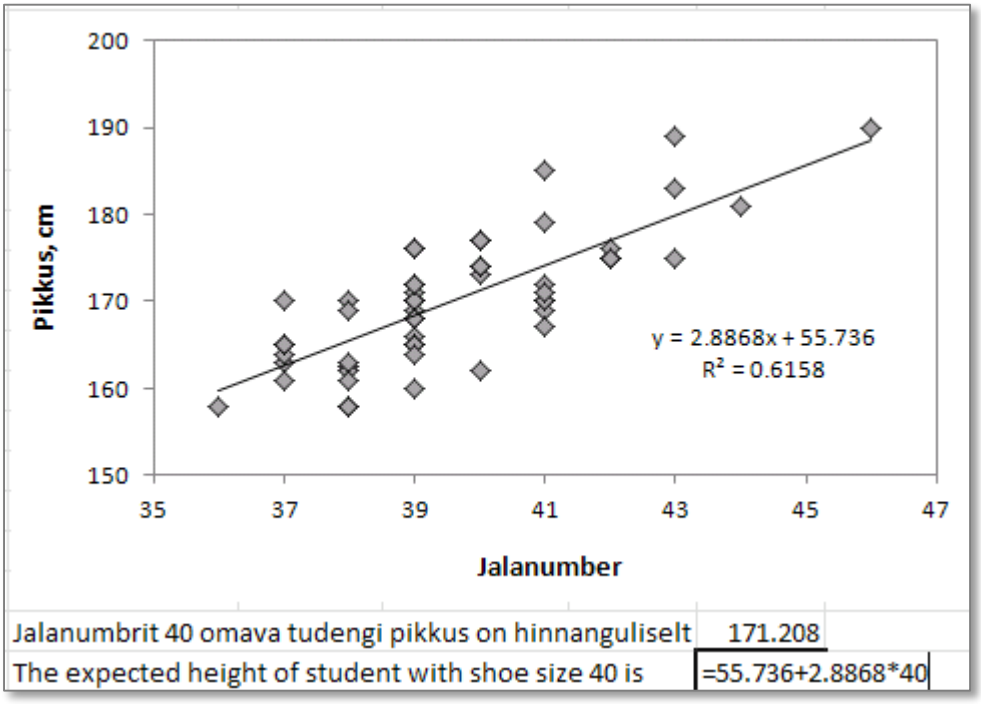


Result:



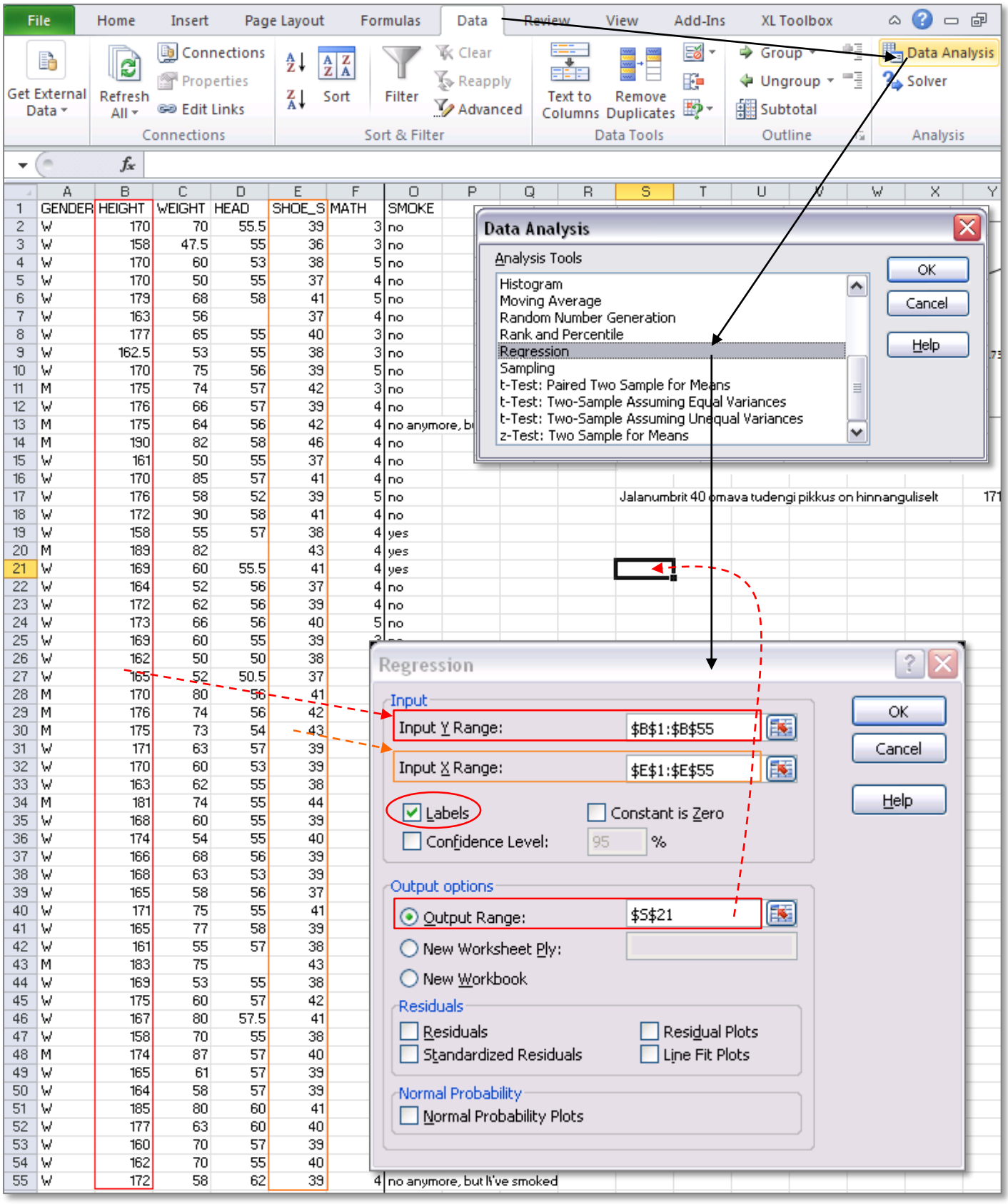
3. Predict based on estimated equation, how tall is expectably the student with shoe size 40.

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…*).



Result of regression analysis:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.7847576
R Square	0.6158445
Adjusted R Square	0.6084569
Standard Error	4.6287793
Observations	54

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1786.077252	1786.077	83.36184	2.18481E-12
Residual	52	1114.131081	21.4256		
Total	53	2900.208333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
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

Height = a + b×Shoe.size,

where instead a and b are their estimated numerical values.

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
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×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					
	df	*SS*	*MS*	*F*	*Significance F*
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

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

predicts students' height more accurately compared with the constant model

$$\text{Height} = a$$

(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	
Regression Statistics	
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), $0 \leqslant R^2 \leqslant 1$. 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 ($R^2 = 0.62$).