Exercise 11

Zoltan Kekecs

8 May 2021

Exercise 11 - The basics of linear regression

This exercise is related to learning the basic logic behind making predictions with linear regression models, and to quantifying the effectiveness of our predictions.

Data management and descriptive statistics

Our dataset

Lets say we work at a shoestore and we would like to be able to tell people's shoe size just by knowing their height. We collect some data using a simple survey about shoe size and height.

We will use the following .sav file in this exercise:

https://github.com/kekecsz/SIMM32/blob/master/2021/Lab_2/Shoesize_data.sav

Check the dataset for irregularities

You should always check the dataset for coding errors or data that does not make sense.

View data in the data editor and display simple descriptive statistics and plots. You can find the commands for data exploration in the **Analyze > Descriptive Statistics tab**, such as:

Analyze > Descriptive Statistics tab > Frequencies

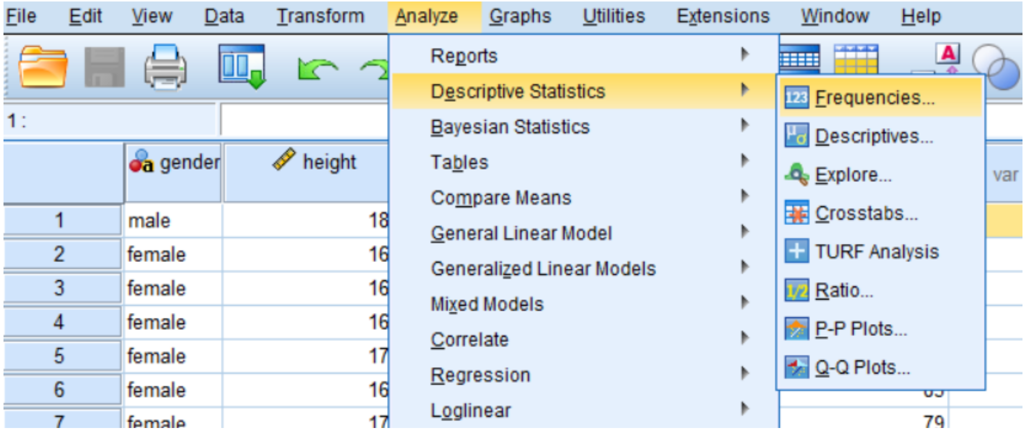
> Frequencies tables will allow you to inspect what kind of values are there in each variable. These values should inform you about whether the data take realistic values.

Analyze > Descriptive Statistics tab > Descriptives

> Descriptives can give you information about the mean and SD, minimum and maximum values, and the skewness and kurtosis of the distribution of the variables (specified in the options when calling this function).

Analyze > Descriptive Statistics tab > Explore

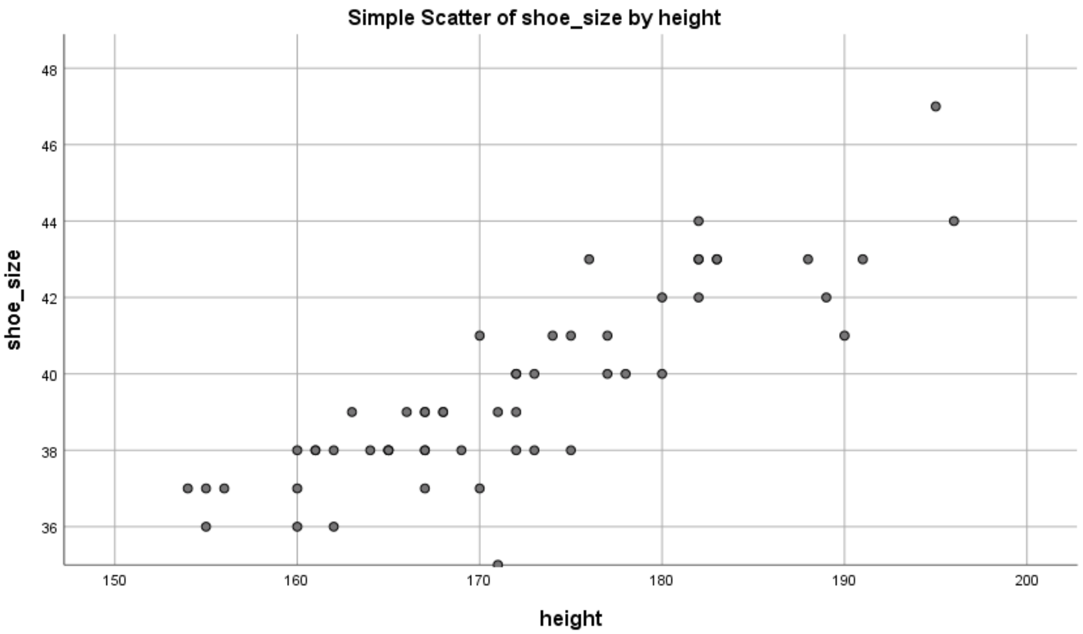
> Explore gives you similar information, but it also includes confidence intervals around the mean, and gives you the option to display a histogram to visually inspect the distribution of the data (specified in the options).



(the text boxes in these documents will show the syntax for the most important manipulations we do in SPSS):

```
*Frequencies
FREQUENCIES VARIABLES=gender height shoesize
  /ORDER=ANALYSIS.


*Describe
DESCRIPTIVES VARIABLES=height shoesize
  /STATISTICS=MEAN STDDEV MIN MAX KURTOSIS SKEWNESS.


*Explore
EXAMINE VARIABLES=height shoesize
  /PLOT BOXPLOT HISTOGRAM NPPLOT
  /COMPARE GROUPS
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

You can also ask for a scatterplot in the **Graphs > Chart builder** by specifying Scatter/Dot from the Galery of possible chart types and dragging shoe_size into the y axis and height into the x axis.



```
*scatterplot


GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=height shoe_size
MISSING=LISTWISE REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

  SOURCE: s=userSource(id("graphdataset"))

  DATA: height=col(source(s), name("height"))

  DATA: shoe_size=col(source(s), name("shoe_size"))

  GUIDE: axis(dim(1), label("height"))

  GUIDE: axis(dim(2), label("shoe_size"))

  ELEMENT: point(position(height*shoe_size))

END GPL.
```

Clean up the dataset and check again to see if everything looks alright now.

Save the cleaned dataset to a new file so that the raw data always remains unchanged!

Prediction with linear regression

how to set up and interpret simple regression

Regression is all about predicting an outcome by knowing the value of predictor variables that are associated with the outcome.

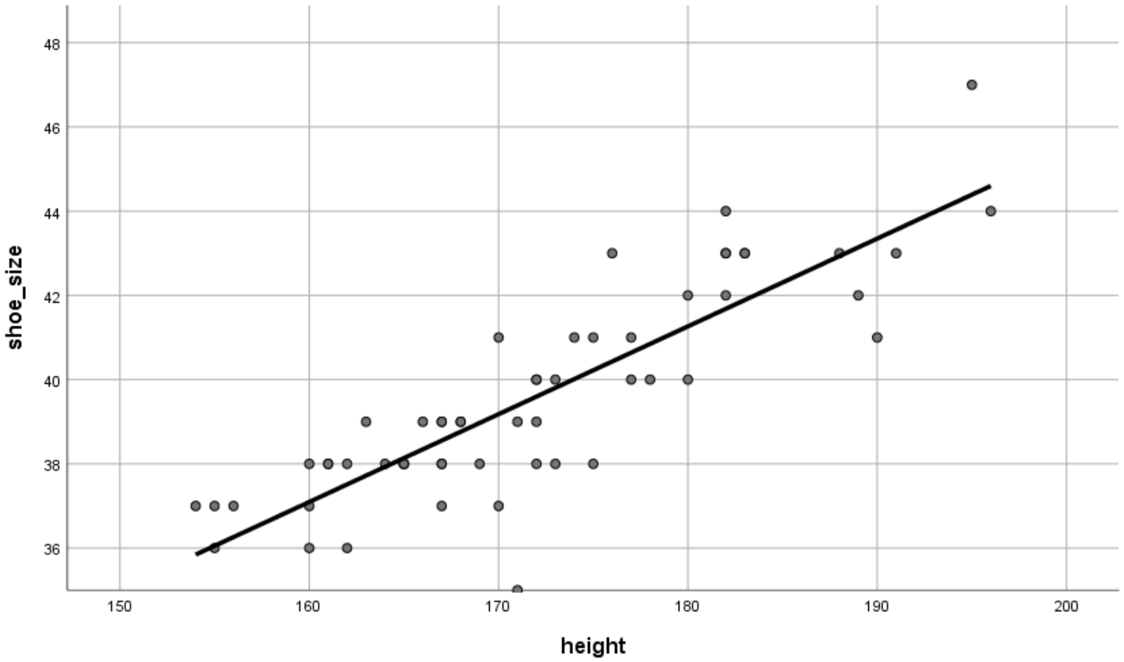
You can set up a simple regression model in the **Analyze > Regression > Linear** tab.

The dependent should be shoesize, because we would like to predict the value of this variable, and the predictor should be height.

```
*simple regression


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT shoe_size
  /METHOD=ENTER height.
```

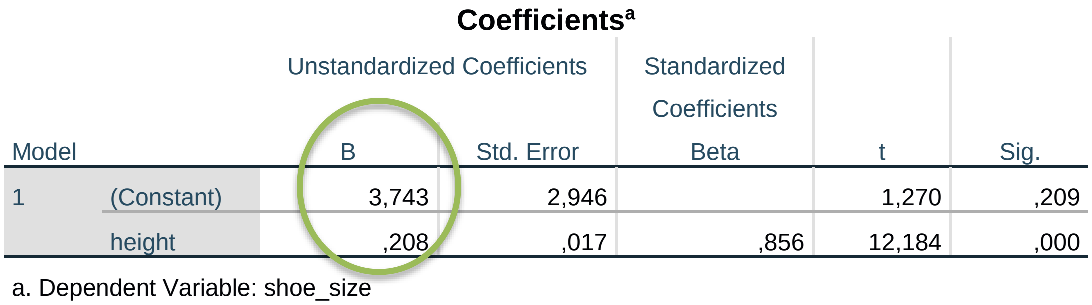
In simple regression, we identify the underlying relationship between the variables, by fitting a single straight line that is closest to all data points.

```
*scatterplot with regression line


GGRAPH

  /GRAPHDATASET NAME="graphdataset" VARIABLES=height shoe_size
MISSING=LISTWISE REPORTMISSING=NO

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

  SOURCE: s=userSource(id("graphdataset"))

  DATA: height=col(source(s), name("height"))

  DATA: shoe_size=col(source(s), name("shoe_size"))

  GUIDE: axis(dim(1), label("height"))

  GUIDE: axis(dim(2), label("shoe_size"))

  ELEMENT: point(position(height*shoe_size))

  ELEMENT: line(position(smooth.linear(height*shoe_size)))

END GPL.
```



Regression provides a mathematical equation called the **regression equation** with which you can predict the outcome by knowing the value of the predictors.

The regression equation is formalized as: Y = b0 + b1*X1, where Y is the predicted value of the outcome, b0 is the intercept, b1 is the regression coefficient for predictor 1, and X1 is the value of predictor 1. You can see this in the Coefficients table in the output of the regression listed as “Unstandardized Coefficients B”.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3,743	2,946		1,270	,209
	height	,208	,017	,856	12,184	,000

a. Dependent Variable: shoe_size

This means that the regression equation for predicting shoe size is:

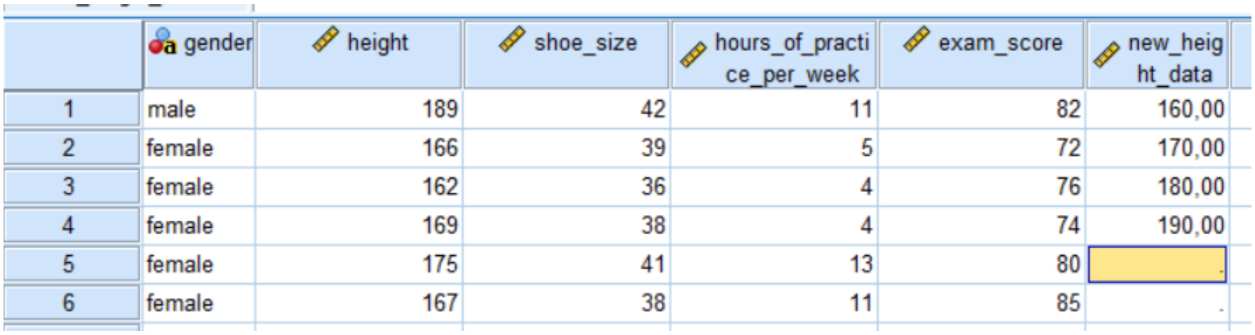
shoe size = 3.37 + 0.21 * height

that is, for a person who is 170 cm tall, the predicted shoe size is calculated this way:

3.37 + 0.21 * 170 = 39.06

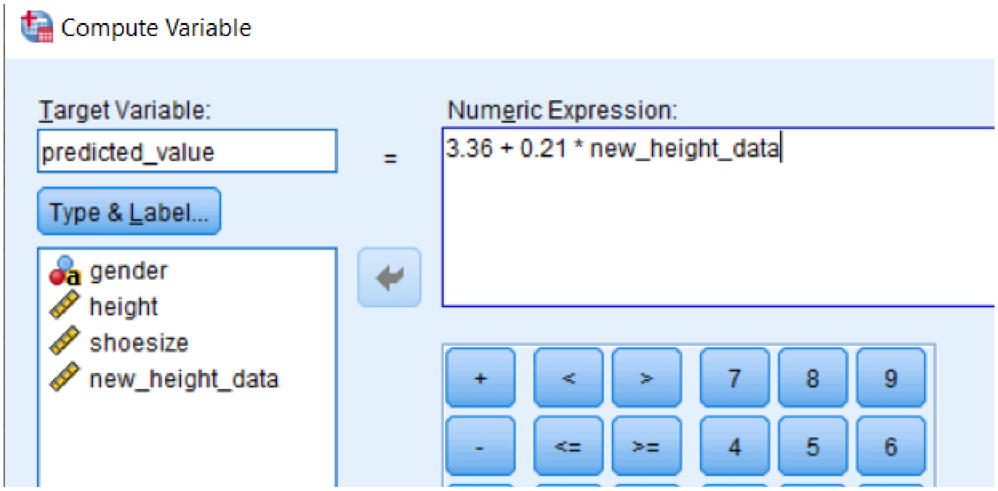
You don’t have to do the calculations by hand, you can ask SPSS to do this for you if you tell it the formula.

For example, we can create a new column (variable) in the variable view. Let’s call it new_height_data. Going back to the data view we can enter specific values here, for which we want to get a predicted shoe size. For example we can enter enter 160, 170, 180, 190 here. It will look something like this:



	gender	height	shoe_size	hours_of_practice_per_week	exam_score	new_height_data
1	male	189	42	11	82	160,00
2	female	166	39	5	72	170,00
3	female	162	36	4	76	180,00
4	female	169	38	4	74	190,00
5	female	175	41	13	80	.
6	female	167	38	11	85	.

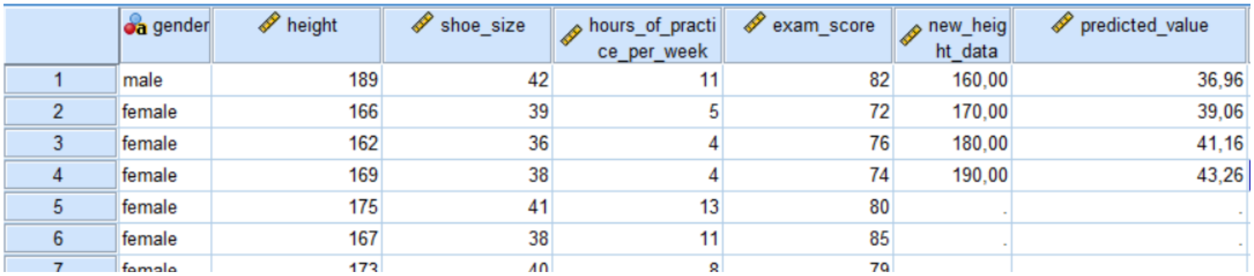
After entering some new data in this new column called new_height_data go to **Transform > Compute variable…** and make spss compute a new variable called “predicted_values” bz specify the regression equation formula:



Target Variable: predicted_value = Numeric Expression: 3.36 + 0.21 * new_height_data

```
*compute predicted value


COMPUTE predicted_value=3.37 + 0.21 * new_height_data.

EXECUTE.
```

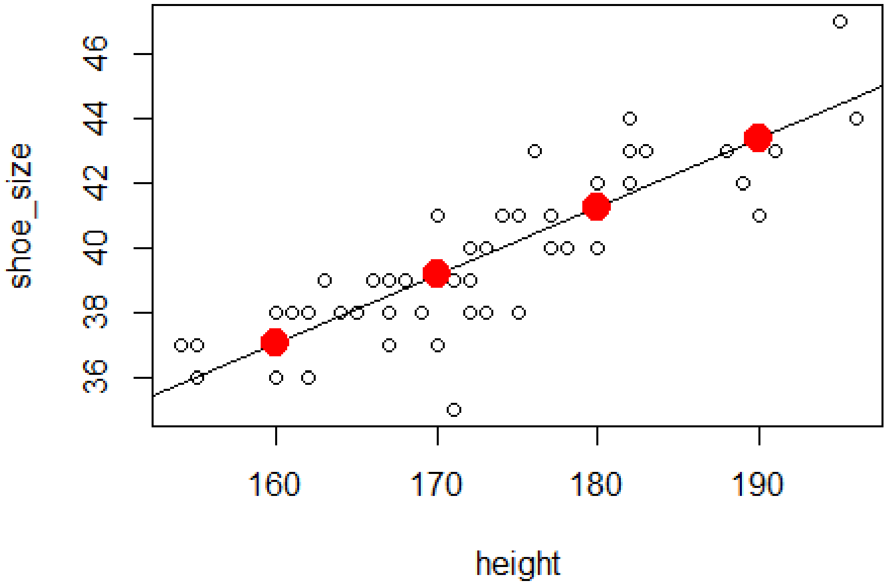
The result is a new column in our dataset that has the predicted shoesize values for each height value we provided.



	gender	height	shoe_size	hours_of_practice_per_week	exam_score	new_height_data	predicted_value
1	male	189	42	11	82	160,00	36,96
2	female	166	39	5	72	170,00	39,06
3	female	162	36	4	76	180,00	41,16
4	female	169	38	4	74	190,00	43,26
5	female	175	41	13	80	.	.
6	female	167	38	11	85	.	.
7	female	173	40	8	79		

Predicted values all fall on the regression line

(this plot was generated using R)



Practice exercise

1. Calculate what is the predicted shoe size for your own height using the procedure we used above!

2. Build a simple linear regression model where **exam_score** is the outcome variable and **hours_of_practice_per_week** is the predictor.

3. Write down the regression equation for estimating exam_score

4. Interpret the regression equation. Do people who practice more have higher or lower exam scores compared to those practicing less?

5. Create a scatterplot of the relationship of exam_score and hours_of_practice_per_week including the regression line and use the plot to verify your interpretation.

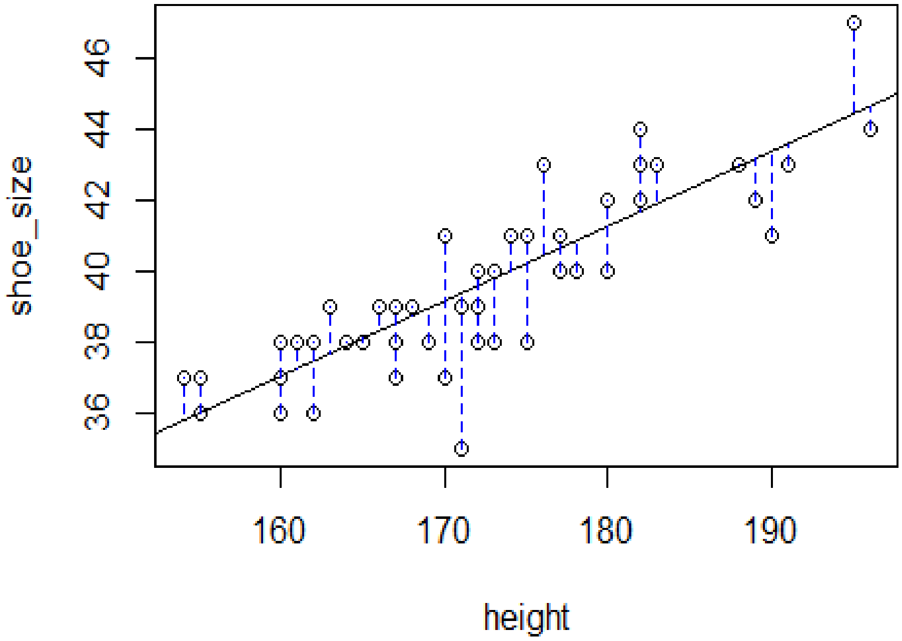
6. Using this model, what is the predicted exam_score for people who practice 2, 5, and 6 hours per week respectively? What is the difference between the person practicing 5 and 6 hours per week, and why?

How good is my model

How to measure prediction efficiency?

You can measure how effective your model is by measuring the difference between the actual outcome values and the predicted values. We call this **residual error** in regression.

The residual error for each observed shoe size can be seen depicted by the blue lines on this plot below (plot generated using R)



You can get all the predicted values for your original data in the **Analyze > Regression > Linear** tab under the **save** button (ask for the unstandardized predicted values). Note that this produces the exact same result as if you would have computed the predicted values using the regression equation in the way we did before.

You can get the residual error by subtracting the predicted value from the actual value of the dependent (predicted) variable. Or you can simply ask for it here: **Analyze > Regression > Linear** tab under the **save** button (ask for the unstandardized residuals). Note that this produces the exact same result as if you would have subtracted the predicted values from the actual observed values of shoe size.

If you ask to save the predicted values and/or residuals, new variables will appear in the data editor, containing these values.

Note that each time you ask for this, new variables will appear!

```
*get predicted values and residuals


REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT shoe_size
 /METHOD=ENTER height
 /SAVE PRED RESID.
```

You can simply add up all the absolute values of the residual error, and get a good measure of the overall efficiency of your model. This is called the residual absolute difference (RAD). However, this value is rarely used. More common is the residual sum of squared differences, or **Residual Sum of Squares (RSS)** for short: take the square of all the residual error scores one-by-one, and add them up.

You can use the **Transform>Compute** function to compute the squared values (saved in a new variable which we can call res_sq for example), and get the sum of these values through using **Analyze>Describe** and specifying in the options that you want the Sum of the variable.

```
*RSS


COMPUTE res_sq=RES_1 * RES_1.

EXECUTE.


DESCRIPTIVES VARIABLES=res_sq

  /STATISTICS=MEAN SUM STDDEV MIN MAX.
```

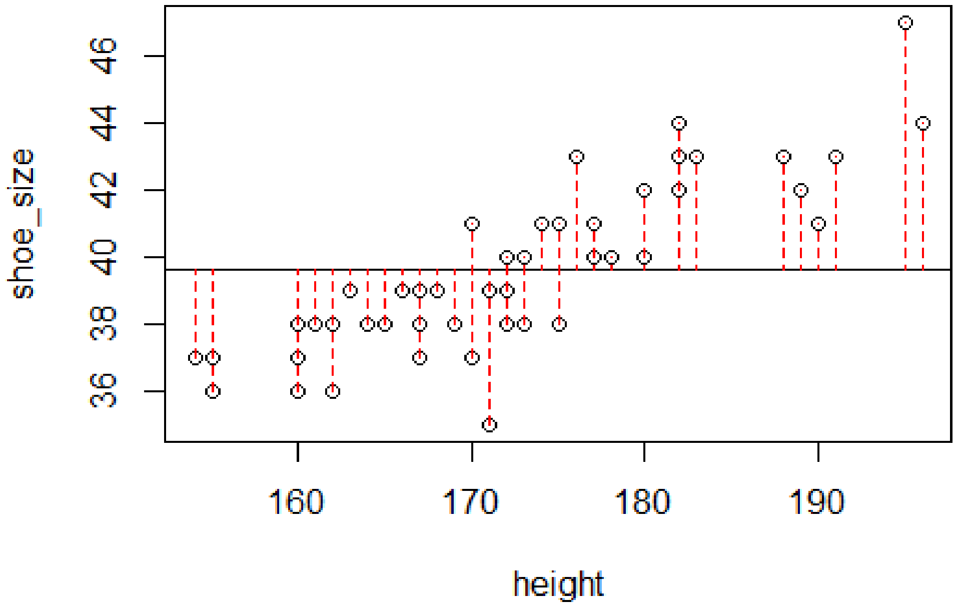
The RSS corresponds to the amount of error that you can expect if you use this particular model for predicting the outcome. So the lower the RSS the better, since a lower RSS means that the predicted values by the model are closer to the actual values that you want to predict.

RSS on its own is not very useful. It is in the raw (squared) units of the outcome variable, and it is a sum of all errors, so it heavily depends on the sample size. The larger the sample the larger the RSS since there is more numbers to add up. So the RSS is not comparable between different samples, or when you would like to compare models with different outcome measures.

Is the predictor useful?

To establish how much benefit did we get by taking into account the predictor, we can compare the residual error when using our best guess (the mean) of the outcome variable's value without taking into account the predictor, with the residual error when the predictor is taken into account.

Below you can find depiction of a regression model where we only use the mean of the outcome variable to predict the outcome value.



We can calculate the sum of squared differences the same way as before, but for the model where we predict with the mean of the outcome variable only, we call this the total sum of squared differences, or **Total Sum of Squares (TSS)** for short.

```
*TSS

DESCRIPTIVES VARIABLES=shoe_size
  /STATISTICS=MEAN STDDEV MIN MAX.

COMPUTE pred_mean=39.571429.
EXECUTE.

COMPUTE res_mean=shoe_size - pred_mean.
EXECUTE.

COMPUTE res_mean_sq=res_mean * res_mean.
EXECUTE.

DESCRIPTIVES VARIABLES=res_mean_sq
  /STATISTICS=MEAN SUM STDDEV MIN MAX.
```

The total amount of information gained about the variability of the outcome is shown by dividing the RSS with the TSS and subtracting it from 1. This statistic is called the **R squared** (R^2). So R^2 = 1 **-** (RSS/TSS). In our case this is 1-(91.15 / 341.71).

You can calculate this by use a calculator (or use the Compute function in SPSS, but it will generate a new variable with the same value for all participants).

```
*R sqared

COMPUTE R_sq = 1-(91.15 / 341.71).
EXECUTE.
```

Fortunately we don’t have to calculate these by hand, since SPSS gives these information in the **model summary** and ANOVA tables in the output of the regression.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,856[a]	,733	,728	1,299

a. Predictors: (Constant), height

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	250,568	1	250,568	148,449	,000[b]
	Residual	91,147	54	1,688		
	Total	341,714	55			

a. Dependent Variable: shoe_size

b. Predictors: (Constant), height

```
*Regression model summary


REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT shoe_size
  /METHOD=ENTER height.
```

This means that by using the regression model, we are able to explain roughly 73% of the variability in the outcome.

R^2 = 1 means all variablility of the outcome is perfectly predicted by the predictor(s)

R^2 = 0 means no variablility of the outcome is predicted by the predictor(s)

Is the model with the predictor significantly better than a model without the predictor?

The anova test will help you find this out, comparing the sum of squares in the two models. This is also shown in the ANOVA table. If the **F-test** is significant (sig < 0.05 for example), it shows that the model with the predictor produces significantly less error (thus, better predictions) than the null model where we only predict with the mean of the outcome.

The **confidence interval** of the regression coefficient can be requested in the **Statistics…** button within **Analyze > Regression > Linear.**

Practice exercise

1. Run the regression analysis again for the model used in the previous practice exercise: where you predicted **exam_score** using **hours_of_practice_per_week** as a predictor.

2. By looking at the output of the model, determine the percentage of variance of exam_score explained when using the model.

3. By looking at the output of the model, determine the RSS and the TSS of the model.

4. By looking at the output of the model, determine whether we use hours of practice as a predictor is significantly better at predicting the exam score than the null model where we only use the mean of exam score in the sample as a predicted value.

5. Compute the RSS of and the TSS of the model through computing the squared residual values, and verify your calculation with the numbers presented in the model summary.