Exercise 11

Zoltan Kekecs

8 May 2020

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 are a company that sells shoes, 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.

The following .sav file lists the collected data:

https://github.com/kekecsz/SIMM32/blob/master/2020/Lab_2/Height%20and%20shoe size.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.

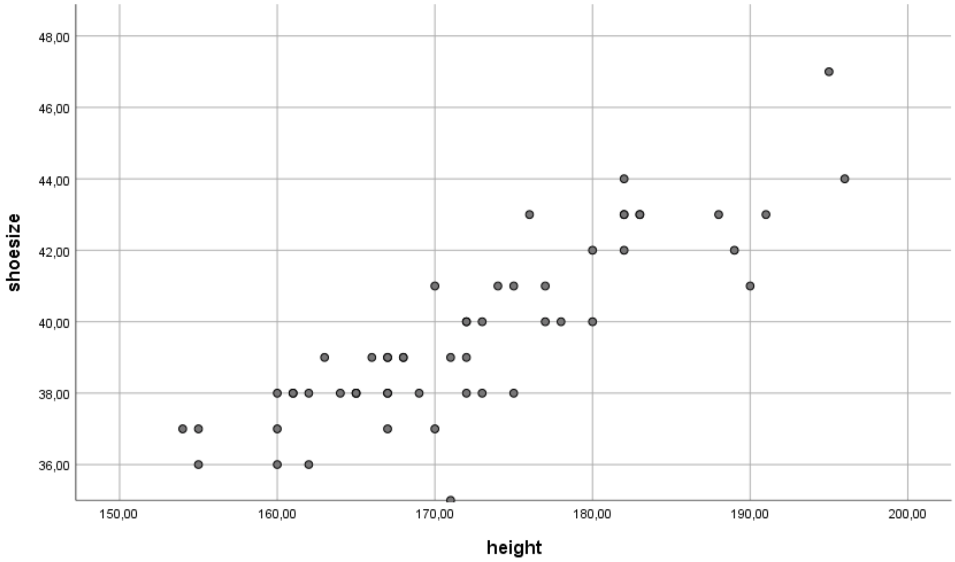
Analyze > Descriptive Statistics tab > Explore

> 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.

(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**



```
*scatterplot


GGRAPH

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

  /GRAPHSPEC SOURCE=INLINE.

BEGIN GPL

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

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

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

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

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

  ELEMENT: point(position(height*shoesize))

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.

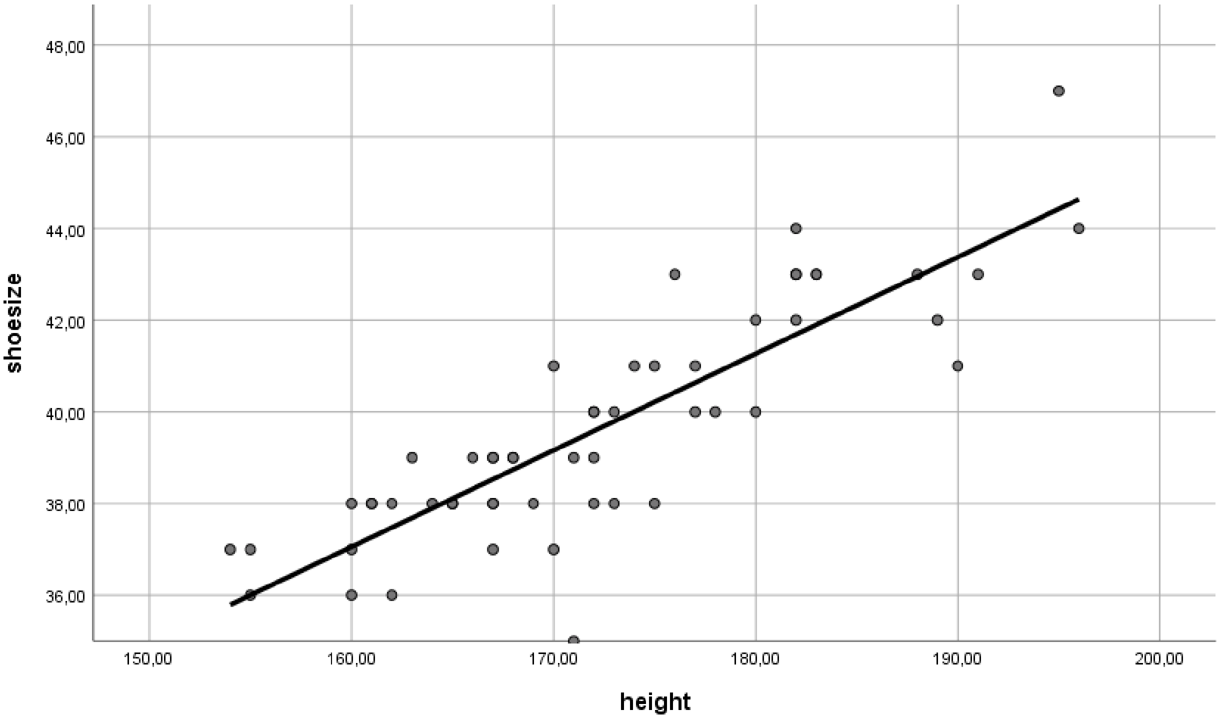
You can set up a simple regression model in the **Analyze > Regression > Linear** tab.

The dependent should be shoesize, because we would like to predict that, and the predictor should be height.

```
*simple regression


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

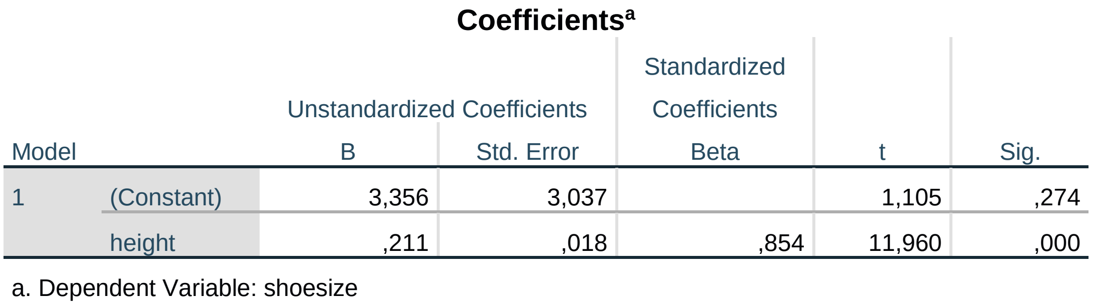
In simple regression, we identify the underlying relationship between the variables, by fitting a single straight line that is closest to all data points.

```
*scatterplot with regression line


GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=height shoesize
MISSING=LISTWISE REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: height=col(source(s), name("height"))
  DATA: shoesize=col(source(s), name("shoesize"))
  GUIDE: axis(dim(1), label("height"))
  GUIDE: axis(dim(2), label("shoesize"))
  ELEMENT: point(position(height*shoesize))
  ELEMENT: line(position(smooth.linear(height*shoesize)))
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)	3,356	3,037		1,105	,274
	height	,211	,018	,854	11,960	,000

a. Dependent Variable: shoesize

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

shoe size = 3.36 + 0.21 * height

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

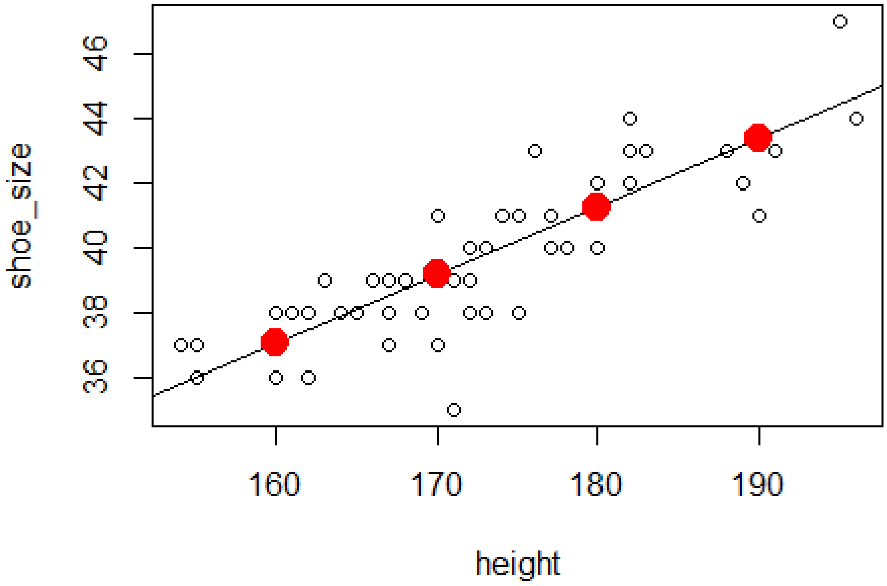
3.36 + 0.21 * 170 = 39.06

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. After entering some new data in a new column we call new_height_datam go to **Transform > Compute variable…** and specify the formula there:

```
*compute predicted value


COMPUTE predicted_value=3.36 + 0.21 * 170.

EXECUTE.
```

Predicted values all fall on the regression line

(this plot was generated using R)

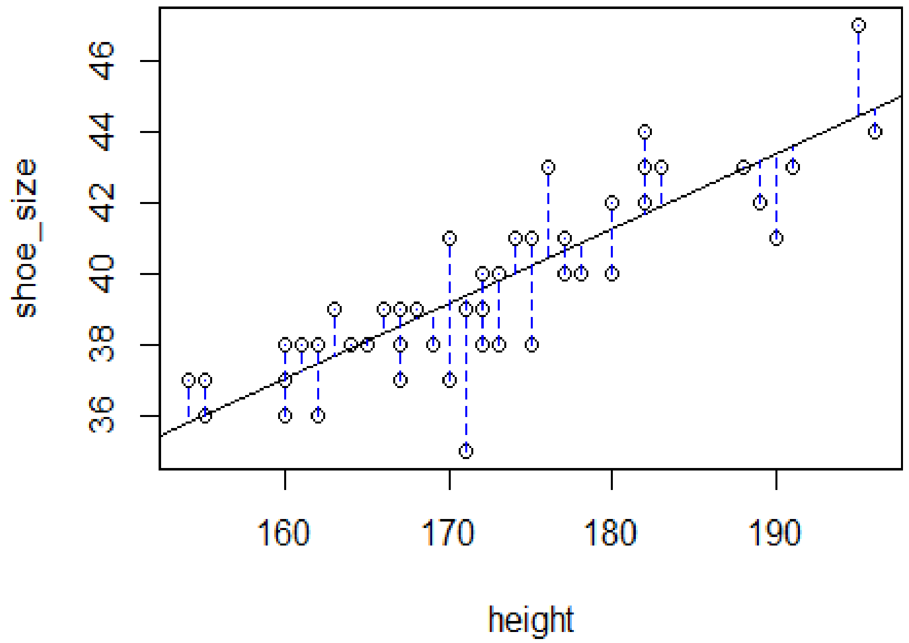


How good is my model

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 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).

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).

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

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 shoesize
 /METHOD=ENTER height
 /SAVE PRED RESID.
```

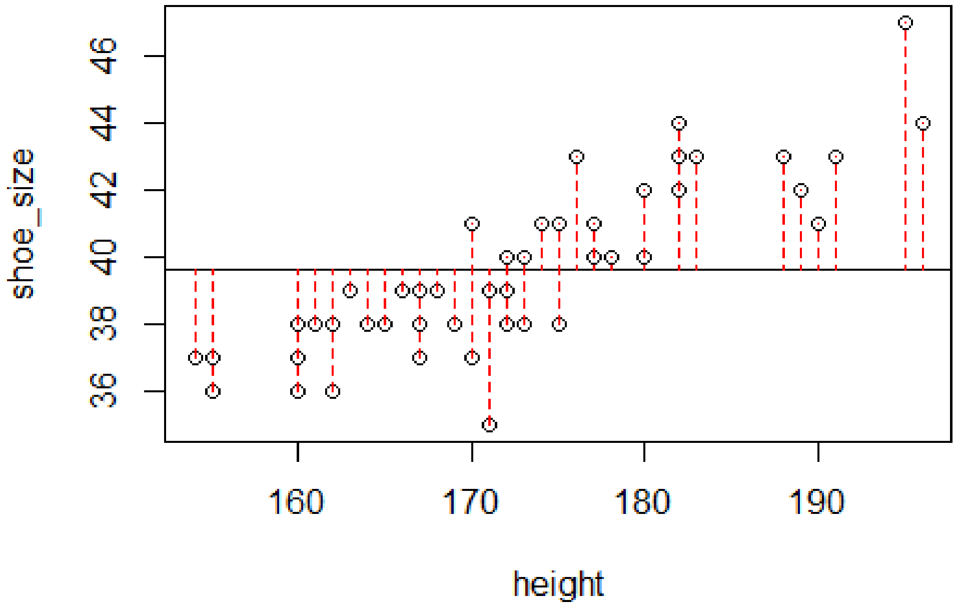
You can simply add up all the absolute values of the residual error, and get a good measure of the overall efficiency of your model. This is called the residual absolute difference (RAD). However, this value is rarely used. More common is the residual sum of squared differences (RSS): take the square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

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 (mean) without taking into account the predictor, with the residual error when the predictor is taken into account.

Below you can find 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 only, we call this the total sum of squared differences (TSS).



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).

Fortunately we don't have to calculate these by hand either, 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	,854[a]	,730	,725	1,30721

a. Predictors: (Constant), height

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	244,415	1	244,415	143,033	,000[b]
	Residual	90,567	53	1,709		
	Total	334,982	54			

a. Dependent Variable: shoesize

b. Predictors: (Constant), 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 error in 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.**