

Excercises 1

Exercise 1

- (a) Two (continuous) variables have a correlation of -0.4 and a covariance of -1.84 . One of the variables has a variance of 4. Calculate the variance of the other variable.
- (b) The transformation between Celsius ($^{\circ}C$) and Fahrenheit ($^{\circ}F$) degrees for temperatures is given by:

$$X_{^{\circ}F} = X_{^{\circ}C} \cdot 1.8 + 32$$

Assume that the average temperature in summer in Berlin is $25^{\circ}C$ where we have a standard deviation of $3^{\circ}C$.

How could you transform these values into $^{\circ}F$? Do also calculate the variances in both degree measures.

Exercise 2

Consider the observations of two variables (quite artificial data! ☺):

x_i	1	1	2	2	3	3
y_i	1	3	2	4	3	5

The following tasks should be solved without R:

- (a) Draw a scatterplot.
- (b) Calculate the regression line.
- (c) Check that the line goes through (\bar{x}, \bar{y}) . (Is that always the case?)
- (d) Calculate the correlation. How do you obtain the coefficient of determination?

Exercise 3

Assume we have observations x_1, \dots, x_n of a variable X . Determine the value a , which minimizes the following criterion:

$$Q(a) = \sum_{i=1}^n (x_i - a)^2$$

Exercise 4

Consider $X \sim N(2, 9)$. What does that mean?

Calculate the following probabilities (with and without R):

- (a) $P(X \leq 0)$ (b) $P(X \leq -1)$
(c) $P(X \geq 5)$ (d) $P(-2 \leq X \leq 2)$

Exercise 5

Consider the following data for the speed versus braking distance (of a car):

Speed X (in km/h)	20	25	30	35	40	45	50	55	60	65	70
Braking distance Y (in m)	18	26	33	40	46	59	72	85	97	120	141

Here are some (possibly) useful values when calculating without R:

$$\bar{x} = 45, \quad \bar{y} = 67, \quad s_X^2 = 275, \quad s_Y^2 = 1600.6, \quad s_{XY} = 649.5$$

- Explain the meaning of: \bar{x} , \bar{y} , s_X^2 , s_Y^2 and s_{XY} .
- What are the R functions to calculate them? (Check with R!)
- Display the data in R using a scatterplot.
- Calculate the correlation and check with R. What could we conclude from this value?
- Calculate the linear regression coefficients (first without R). Do the same using R and display the line in the scatterplot.

Exercise 6

On Moodle you find the file `Miete2003.csv`. This is a sample of 2053 apartments in Munich from 2003 (Munich was already an expensive city at this time ...). The variables are coded as follows:

Variable	Meaning	Values
<code>nettomiete</code>	net rent (per month)	in Euro
<code>nettomiete.qm</code>	net rent per square metre	in Euro
<code>wohnflaeche</code>	living space	in square metres
<code>zimmerzahl</code>	no. of rooms	1,...,6
<code>baujahr</code>	year of construction	year
<code>bezirk</code>	Munich district	1,...,25
<code>wohnlage</code>	quality of location	beste (best), einfache (simple), gute (good)
<code>warmwasser</code>	warm water provided	ja (yes), nein (no)
<code>z.heizung</code>	central heating	ja (yes), nein (no)
<code>bad.kacheln</code>	bath room with tiles	ja (yes), nein (no)
<code>bad.extras</code>	bath room with extras	ja (yes), nein (no)
<code>geh.kueche</code>	kitchen with upscale equipment	ja (yes), nein (no)

Original source: <https://doi.org/10.5282/ubm/data.2>

- For which pairs of variables would it be useful to estimate a simple linear regression model? (Choose at least two different examples. Explain which of the variables do you consider the dependent and the independent one.)
- Load the data into R. If you don't know what to do, check `?read.csv2`. (Extra task: Try also to load the original data into R!)

- (c) Estimate the model that you have chosen in (a), i.e. calculate the coefficients, draw scatterplots and regression lines, determine R^2 .
- (d) Now consider `nettomiete` and `wohnlage`. Do you think it is useful to consider simple linear regression here? Do you know any other technique(s) to analyse the relationship between these two? (Maybe also a graphical technique?)
- (e) Again consider two variables: `bad.kacheln` and `geh.kueche`. Do you think there is a relationship between these two? How could you analyse it?