Regression

Master Data Science
Summer 2018

Prof. Dr. Marlene Müller
marlene.mueller@beuth-hochschule.de

Exercises 1 (incl. hints to solve)

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.

$$r_{XY} \quad = \quad -0.4 = \frac{s_{XY}}{s_X \cdot s_Y} = \frac{-1.84}{s_X \cdot \sqrt{4}} \quad \Rightarrow \quad s_X = 2.3$$

(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.

$$\overline{x}_F = \overline{x}_C \cdot 1.8 + 32 = 77\,, \quad s_F^2 = s_C^2 \cdot 1.8^2 = 9.72 \quad \Rightarrow \quad s_F = s_C \cdot 1.8 = 5.4$$

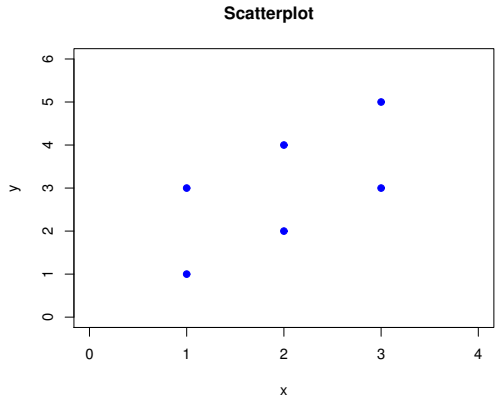
Exercise 2

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.
$\widehat{\beta}_0 = 1, \widehat{\beta}_1 = 1 \;\Rightarrow\; \widehat{y} = 1 + x$

(c) Check that the line goes through $(\overline{x}, \overline{y})$. (Is that always the case?)
$\overline{x} = 2\,, \overline{y} = 3 \;\Rightarrow$ yes! (as $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$)

(d) Calculate the correlation. How do you obtain the coefficient of determination?
$r_{XY} = 0.6324555 \;\Rightarrow\; R^2 = r_{XY}^2 = 0.4$

Exercise 3

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

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

$$\begin{aligned} Q(a) &= \sum_{i=1}^{n} (x_i - a)^2 = \sum x_i^2 - 2a \sum x_i + na^2 \\ \frac{dQ}{da} &= -2 \sum x_i + 2na \stackrel{!}{=} 0 \quad \Rightarrow \quad a = \frac{\sum x_i}{n} = \overline{x} \\ \frac{d^2Q}{da^2} &= 2n > 0 \quad \text{(minimum)} \end{aligned}$$

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) $P(-2 \leq X \leq 2)$

$$\begin{aligned} P(X \leq 0) &= P\left(\frac{X-2}{3} \leq \frac{0-2}{3}\right) = \Phi\left(-\frac{2}{3}\right) \approx \Phi(-0.67) \approx 0.25143 \\ P(X \leq -1) &= P\left(\frac{X-2}{3} \leq -1\right) = \Phi(-1) \approx 0.15866 \\ P(X \geq 5) &= 1 - P(X \leq 5) = 1 - \Phi(1) \approx 0.15866 \\ P(-2 \leq X \leq 2) &= P\left(-\frac{4}{3} \leq \frac{X-2}{3} \leq 0\right) = \Phi(0) - \Phi\left(-\frac{4}{3}\right) \approx 0.40824 \end{aligned}$$

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:

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

(a) Explain the meaning of: $\overline{x}$, $\overline{y}$, s_X^2, s_Y^2 and s_{XY} .
means of X and Y, variances of X and Y, covariance

(b) What are the R functions to calculate them? (Check with R!)
`mean`, `var`, `cov`

(c) Display the data in R using a scatterplot.

(d) Calculate the correlation and check with R. What could we conclude from this value?

$$r_{rx} = \frac{s_{XY}}{s_X \cdot s_Y} \approx 0.9789746$$

(e) Calculate the linear regression coefficients (first without R). Do the same using R and display the line in the scatterplot.

$$\widehat{\beta}_1 = \frac{s_{XY}}{s_X^2} \approx 2.362\,, \quad \widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \cdot \overline{x} \approx -39.282$$

$\Rightarrow$ `Exercises1-5.R`

Exercise 6

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

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

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

(a) 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.)

(b) 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?

⇒ `Exercises1-6.R`