Statistical Computing

Master Data Science

Winter Semester 2017/18

BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 7 — Regression

Section 2 includes an exercise, to be done before next week's workshop.

1 In the Workshop

Preliminaries

- Download the data file `Nebenwirkung.Rda` from Moodle page into your `H:\\StatComp` directory.
- Start RStudio
- Strg+Shift+N opens a new R script.
- Type in the comments

  ```
  #Statistical Computing: Workshop 6
  #Regression
  ```

- Save the file in your `H:\\StatComp` folder with the name `Workshop6.R`.
- Set your *working directory* to be `H:\\StatComp`. The code to do this is

  ```
  > setwd("H://StatComp")
  ```

- Clear your workspace using of objects from a previous session *Session > clear workspace*.
- Open a Word document or similar to answer the exercises in this workshop.

Exercise 1 Regression coefficients

In this exercise, you will calculate coefficients using the formulae given in the lecture.

In order to see how the number of guests in a hotel affects water consumption, a hotel manager collected data on the hotel's water consumption and the hotel occupancy (number nights' stay for all guests) over $n = 5$ weeks.

i	1	2	3	4	5
Occupancy x_i	20	50	70	100	100
Water consumption y_i	25	35	20	30	45

(a) Define two R Objects `occupancy` und `consumption` using the above data.

(b) Plot the two variables in a scatter plot.

(c) Calculate the following statistics (entering the answers in the Word document).

(i) $\overline{x}$

(ii) $\overline{y}$

(iii) $\sum_{i=1}^{5}(x_i - \overline{x})^2$

(iv) $\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y})$

(v) s_x^2

(vi) s_{xy}

(vii) Gradient $\widehat{b}_1$ und

(viii) Intercept $\widehat{b}_0$

(d) Write down the regression function

(e) Add the regression line to the scatter plot. [`abline(c(a,b))` draws a line with intercept `a` and gradient `b`]

(f) What is the water consumption according to the regression model when the hotel has an occupancy of 70 guest-nights?

(g) Calculate the 5 residuals.

Exercise 2 Regression using `lm`

You will now repeat Exercise 1 but using the normal commands to fit a simple linear regression in R using the command `lm(y~x)` or `lm(y~x,data=dataframe)`. The second version is used when `x` and `y` are variables in `dataframe`. At each stage check that your results match up to those in Exercise 1.

(a) Fit the linear regression model to the hotel data, and assign the result to the object called `lm.obj1`:

```
> lm.obj1<-lm(consumption~occupancy).
```

(b) Look at the results:

```
> summary(lm.obj1)
```

(c) find the following in the output

   (i) b_0

   (ii) b_1

   (iii) The p-value of the variable `occupancy`

   (iv) "R^2"The "goodness of fit" statistic.

   N.B. The R^2 is small and the p-value is large. This suggests that water consumption is not dependent on the level of occupancy. As we would expect there to be a dependency, it is likely that because $n = 5$ there is insufficient data to show this effect.

(d) Output the fitted values

```
> fitted(lm.obj1)
```

(e) What is the water consumption according to the regression model when the hotel has an occupancy of 70 guest-nights?

(f) Output the five residuals

```
> resid(lm.obj1)
```

(g) Obtain a residual plot for this model:

```
> plot(lm.obj1,which=1)
```

There are not enough data points to properly asses this model.

Exercise 3 Side effects
A new drug to treat depression is suspected to have the side effect of delayed reactions. Ten randomly chosen patients were given varying doses of the drug. The reaction times of the patients were measured in an experiment (larger values mean slower reactions). The resulting data can be found in the datafile you downloaded `Nebenwirkung.rda`

(a) Load the data:

```
> load("Nebenwirkung.rda")
```

(b) The name of the data frame and variable names are given in German! rename them!

```
> library(dplyr)
> sideeffect<-rename(sideeffect,dose=Dosierung, reaction=Reaktion)
> rm("Nebenw")
```

(c) Obtain the six-number summary for each variable in the data frame

```
> summary(sideeffect)
```

(d) Plot the data in a scatterplot (reaction time **against** dose ⇒ y axis **against** x axis).

(e) Before you compute the regression coefficients, try to "guestimate" from the graph approximate values for the intercept and gradient.

(f) Fit a linear regression model with `reaction` as the dependant variable and `dose` as the dependent variable.

(g) Output the summary of the regression model.

(h) What are the values of $\widehat{b}_0$ and $\widehat{b}_1$? Write down the regression formula.

(i) Does dosage have a statistically significant effect on reaction time?

2 Tidying up

► Tidy up your script file including sensible comments.

► Save the script file (source file) again: $\boxed{\text{Strg + S}}$ or *File > Save as*.

► Leave RStudio by typing the command:

```
> q()
```

When R asks you *Save workspace image ...?*, click on **Don't save**!

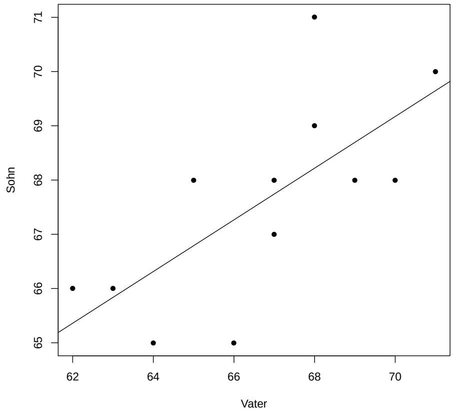
► Feierabend!

3 Homework exercise

The table below contains the heights of fathers x and their sons y. The data are from an american study so are given in inches (1 inch = 2.54 cm).

Do the heights of the ‘sons depend on the heights of their fathers?

Fit a simple linear regression of the form $y_i = \widehat{b}_0 + \widehat{b}_1 x_i + \epsilon_i$ to the 12 father-son pairs.



Father	Son						
x	y	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})(y_i - \overline{y})$	$\widehat{y}_i$	$y_i - \widehat{y}_i$
65	68	-1.67	0.42	2.78	-0.69		
63	66	-3.67	-1.58	13.44	5.81	65.84	0.16
67	68	0.33	0.42	0.11	0.14	67.74	0.26
64	65	-2.67	-2.58	7.11	6.89	66.31	-1.31
68	69	1.33	1.42	1.78	1.89	68.22	0.78
62	66	-4.67	-1.58	21.78	7.39	65.36	0.64
70	68	3.33	0.42	11.11	1.39	69.17	-1.17
66	65	-0.67	-2.58	0.44	1.72	67.27	-2.27
68	71	1.33	3.42	1.78	4.56	68.22	2.78
67	67	0.33	-0.58	0.11	-0.19	67.74	-0.74
69	68	2.33	0.42	5.44	0.97	68.69	-0.69
71	70	4.33	2.42	18.78	10.47	69.65	0.35
Totals 800	811	0	0	84.67	40.33		

The extra columns have been provided to make the calculations less time consuming.

(a) Calculate the following

(i) $\bar{x}$

(ii) $\bar{y}$

(iii) the variance of x

(iv) the covariance of x and y

(b) Determine the regression coefficients $\widehat{b}_1$, $\widehat{b}_0$, and give the formula for the regression line.

(c) Calculate the first fitted value $\widehat{y}_1$.

(d) Calculate the first residual $\widehat{\epsilon}_1$.

(e) Show that the regression line passes through the point $(\bar{x}, \bar{y})$