Statistical Computing
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 6 — Regression 1

Section 2 includes an exercise, to be done before next Friday.

1 During the Workshop

Preliminaries

- Start RStudio, and start a new R script `Ctrl+Shift+N`.
- 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 use R to calculate coefficients using the formulae given in the lecture, to gain a better understanding of the calculations involved in fitting a regression model.

In order to see how the number of guests in a hotel affects water consumption, a hotel manager collected weekly data on the hotel's water consumption (Thousand litres per guest per night) and the hotel occupancy (number of guest-nights) 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` and `consumption` using the above data. Which of the two variables corresponds to x, in the classical regression notation, and which variable corresponds to y? *`occupancy` is x and `consumption` is y*

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

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

(i) $\overline{x} = 68$

(ii) $\overline{y} = 31$

(iii) $\sum_{i=1}^{5}(x_i - \overline{x})^2 = 4680$

(iv) $\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) = 610$

(v) $s_x^2 = 1170$

(vi) $s_{xy} = 152.5$

(vii) gradient $\widehat{b}_1$ (hint: see lecture notes) $= 0.1303$ and

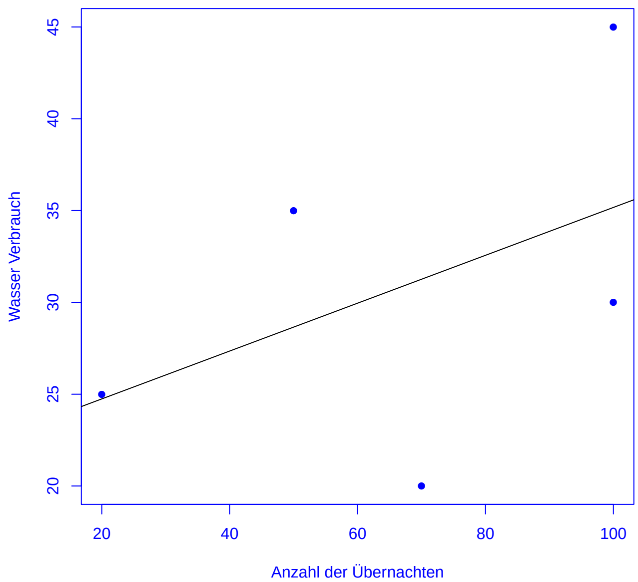
(viii) intercept $\widehat{b}_0 = 22.137$

(d) Write down the regression function $f(x) = 22.137 + 0.1303x$

(e) Add the regression line to the scatter plot. Hint: `abline(c(a,b))` draws a line on the existing plot with intercept `a` and gradient `b`] *see below*

(f) What is the water consumption according to the regression model when the hotel has an occupancy of 70 guest-nights? This is called the predicted value. $= 31.268$

(g) Calculate the 5 residuals. *0.2564, 6.346, -11.26, -5.171, and 9.829*



Exercise 2 Regression using `lm`

You will now repeat Exercise 1 but using the usual R commands to fit a simple linear regression 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 $\widehat{b}_0$ and $\widehat{b}_1$ in the output.

(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? $= 31.268$

(f) Output the five residuals
`> resid(lm.obj1)`

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**!

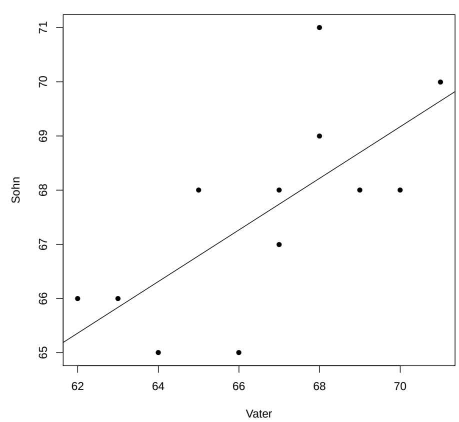
► Feierabend!

2 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 $y_i = \widehat{b}_0 + \widehat{b}_1 x_i + \widehat{\epsilon_i}$ to the 12 father-son pairs.



Father x	Son 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) $\bar{x}$ $\bar{x} = 66.67$

(ii) $\bar{y}$ $\bar{y} = 67.583$

(iii) the variance of x $s_x^2 = 7.697$

(iv) the covariance of x and y $s_{xy} = \dfrac{40.33}{11} = 3.667$

(b) Determine the regression coefficients $\widehat{b}_1$, $\widehat{b}_0$, and give the formula for the regression line. $\widehat{b}_1 = \dfrac{40.33}{84.67} = \dfrac{3.667}{7.696} = 0.4764$, $\widehat{b}_0 = 67.58 - 0.48 \times 66.67 = 35.8248$
$y = 35.8248 + 0.4764x$

(c) Calculate the first fitted value $\widehat{y}_1$ (missing from the table). *66.79*

(d) Calculate the first residual $\widehat{\epsilon}_1$ (missing from the table). *1.21*

(e) Show that the regression line passes through the point $(\bar{x}, \bar{y})$ $\bar{x} = 66.67$ *Regression formula gives* $y = 35.82 + 0.48\bar{x} = 67.5848$ *compare with* $\bar{y} = 67.583$. *The difference is just numerical rounding error.*