Prof. Dr. Steffen Wagner
Angewandte Statistik
Fachbereich II
Berliner Hochschule für Technik

Exercise 06: Simple linear regression

Statistical Computing – WiSe 2022/23

Please start this exercise with the written exercise to recap the basics of simple linear regression. Afterwards continue with the R exercises.

Written exercise

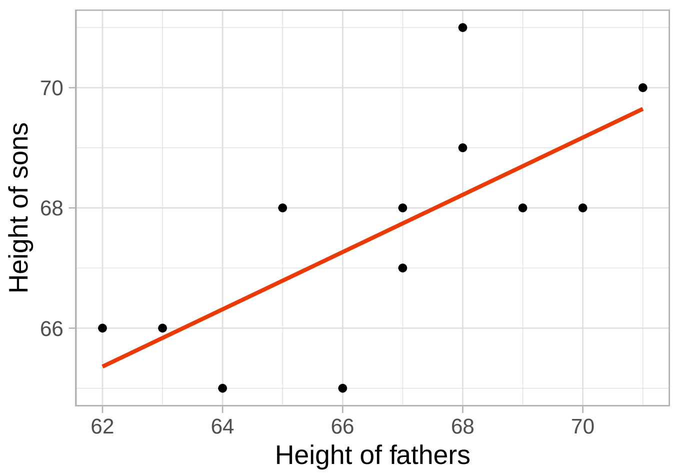
The following table[1] gives the heights of fathers x and their sons y. The data are from an American study so are given in inches (1 inch = 2.54 cm).

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$	$\widehat{y}_i$	$y_i - \widehat{y}_i$
	65	68	-1.67	0.42	2.78	-0.69	66.79	1.21
	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	84.70	40.40	811	0

Question: Do the heights of the sons depend on the heights of their respective fathers?

To answer this question, 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 using the following steps. The scatterplot and regression line is shown below.

a) *Guesstimate* the slope of the regression line from the scatter plot. A guessed slope would be $\widehat{b}_1^{eye} = 0.5$



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

a) Calculate the following:
- $\overline{x} = 66.667$
- $\overline{y} = 67.583$
- the variance of x $s_x^2 = 7.7$
- the covariance of x and y $s_{xy} = \frac{40.4}{11} = 3.673$

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

$$\widehat{b}_1 = \frac{40.4}{84.7} = \frac{3.673}{7.7} = 0.476 \quad \text{and} \quad \widehat{b}_0 = 67.583 - 0.476 \cdot 66.667 = 35.825 \quad \Rightarrow \widehat{y} = 35.825 + 0.476x$$

c) Calculate the first fitted value $\widehat{y}_1$ (missing from the table). 66.789

d) Calculate the first residual $\widehat{\epsilon}_1$ (missing from the table). 1.211

e) Show that the regression line passes through the point $(\overline{x}, \overline{y})$
With $\overline{x} = 66.667$ Regression formula gives $\widehat{y} = 35.825 + 0.476 \cdot 66.667 = 67.558$ compare with $\overline{y} = 67.583$. The difference is just numerical rounding error.

The R exercises today have again fewer commands for you to blindly type in, than in previous weeks. For aspects covered in previous workshops/lectures consult the relevant teaching material. In other places hints are given. At the end there is a written exercise for you to practice calculating the correlation.

R exercises

Preparations

a. Make sure you work inside your RStudio project. Details please see Exercise 01.

b. Open a new R script file (either `Shift + Ctrl + N` or via the menu), type the following comment/code in the first few lines and save the file.

```
#==========================================================================#
# Statistical Computing: exercise sheet 6                                  #
# Simple linear regression                                                 #
#==========================================================================#
rm(???)
```

c. Fill in the missing argument `???` to call of the function `rm`.

Regression coefficient

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.

	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

```
occupancy <- c(20, 50, 70, 100, 100)
consumption <- c(25, 35, 20, 30, 45)
```

b) Plot the two variables in a scatter plot and label the axis. Add a plot title.

```
plot(occupancy, consumption,
     xlab = "Occupancy x",
     ylab = "Water consumption",
     main = "Hotel Data")
```

c) Calculate the following statistics and save the results as a comment in your R code.

- $\overline{x} = 68$
- $\overline{y} = 31$
- $\sum_{i=1}^{5}(x_i - \overline{x})^2 = 4680$

- $\sum_{i=1}^{5}(y_i - \overline{y})^2 = 370$
- $\sum_{i=1}^{5}(x_i - \overline{x})(y_i - \overline{y}) = 610$
- $s_x^2 = 1170$
- $s_y^2 = 92.5$
- $s_{xy} = 152.5$
- slope $\widehat{b}_1$ (hint: see lecture notes) $= 0.1303$
- intercept $\widehat{b}_0$ $\widehat{f}(75) = 22.1368$

d) Write down the regression function. $f(x) = 22.1368 + 0.1303x$

e) Add the regression line to the scatter plot using the function `abline`. Hint: read the help for `abline`.

```
b0 <- ???   # calculate the intercept
b1 <- ???   # calculate the slope
abline(coef = c(b0, b1))
```

f) What is the water consumption according to the regression model when the hotel has an occupancy of 75 guest-nights? This is called the predicted value. $= 31.9093$

g) Calculate the 5 residuals r_i and their sum $\sum_i r_i$.
$r_1 = 0.2564$, $r_2 = 6.3462$, $r_3 = -11.2607$, $r_4 = -5.1709$, $r_5 = 9.8291$ and $\sum_i r_i = 0$.

h) What proportion of the variance of the observed water consumption can be explained by the occupancy? Calculate the Pearson correlation coefficient for water consumption and occupancy.

$$R^2 = \frac{\widehat{b}_1 \cdot s_X^2}{s_Y^2} = \frac{0.1303 \cdot 152.5}{92.5}, = 0.2148 \quad \Rightarrow \quad r_{X,Y} = \sqrt{R^2} = 0.4635$$

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 columns 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 `lm01`.

```
lm01 <- lm(consumption ~ occupancy)
```

b) Run the following code and describe the outcome.

```
lm01
coef(lm01)
summary(lm01)
fitted(lm01)
resid(lm01)
```

c) Predict the water consumption when the hotel has an occupancy of 75 guest-nights. Make use of the function `predict()`. Check the help for `predict.lm` to understand, what object the argument `newdata` expects.

```
# new data.frame with x values saved in a variable with identical name as x
# in the specified model formula, here: consumption ~ occupancy
predDf <- data.frame(occupancy = 75)
```

```
predict(lm01, newdata = predDf)
```