Regression

Master Data Science
Summer 2023

Prof. Dr. Marlene Müller
marlene.mueller@bht-berlin.de

Exercises 2

Exercise 1

Consider the following data on prices (in Euro) and sales:

	1	2	3	4	5	6	7	8	9	10
Sales	1585	1819	1647	1496	921	1278	1810	1987	1612	1413
Price	12.50	10.00	9.95	11.50	12.00	10.00	8.00	9.00	9.50	12.50

Economic theory assumes a relationship between both.

(a) Plot the data. How would you describe this relationship? What sign of the correlation do you expect?

(b) Calculate (using R) the estimated coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ for a simple linear regression to model sales in dependence of prices. Add the regression line to your scatterplot.

(c) Is there any connection between r_{XY} and $\widehat{\beta}_1$?

(d) Using the estimated linear model: Predict the sales for a price of 13 Euro.

(e) Discuss: What happens to the estimated coefficients if the prices are converted into another currency? How do the the correlation and the coefficient of determination change?

(You find exchange rates for different currencies under:
http://www.oanda.com/currency/converter/)

Functions in R

Optional input parameters can be omitted, see the following code examples. To give back more than one output, lists are used.

Try out the following code examples:

```
myfun1 <- function(x, a){
  r <- a*sin(x)
  return(r)
}
myfun1(pi/2,2)

## or equivalently:
myfun1 <- function(x, a) a*sin(x)
```

```
myfun2 <- function(x, a=1){       ## a is optional with default 1
  a*sin(x)
}
myfun2(pi/2,2)
myfun2(pi/2)

myfun3 <- function(x, a=NULL){    ## a is optional without default
  if (!is.null(a)){ a*sin(x) }else{ cos(x) }
}
myfun3(pi/2,2)
myfun3(pi/2)

myfun4 <- function(x, a=1){
  r1 <- a*sin(x); r2 <- a*cos(x)
  return(list(r1=r1,r2=r2))       ## list of 2 outputs
}
myfun4(pi/2)
myfun4(pi)$r1


myfun5 <- function(x, a=1, b=2){
  r1 <- a*sin(x); r2 <- b*cos(x)
  return(list(r1=r1,r2=r2))
}
myfun5(pi/2)            ## a=1, b=2 (defaults)
myfun5(pi/2,1,2)        ## a=1, b=2 (both are specified)
myfun5(pi/2,2)          ## a=2, b=2 (only a is specified)
myfun5(pi/2,a=2)        ## a=2, b=2 (only a is specified)
myfun5(pi/2,b=3)        ## a=1, b=3 (only b is specified)
```

Exercise 2

Recall the estimates in simple linear regression:

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

(a) Write your own R function `simple.lm`, that calculates these coefficients and returns the vector $\widehat{\beta} = \left(\widehat{\beta}_0, \widehat{\beta}_1\right)^\top$.
(The input of this function should be the two data vectors `x` and `y`.)

(b) Apply your R function to a simple regression example from the Munich appartments data set (`MunichRent2003.csv`).

(c) Add the calculation of R^2 to your function. (See the formula on slide 13 for example.)

(d) Compare your results with that of the R built-in function `lm`.

Exercise 3

(a) Write a R function that calculates the polynomial

$$p(x) = c_0 + c_1 x + c_2 x^2 + \ldots + c_p x^p$$

where the given input is the vector $(c_0, c_1, c_2, \ldots, c_p)$. The polynomial shall be calculated on a grid of x-values in $[-5, 5]$.

(b) Add an optional way to let the user modify the interval for the x-values.

(c) Add another option to graph the function. The user should also have the possibility to change the color, linestyle and title of the graph.

Exercise 4

Generate artificial regression data using the follwing setup: $X \sim N(0,1),\ \varepsilon \sim N(0, \sigma^2)$ where $\sigma^2 = 0.49$ and

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Choose two values for β_0 and β_1 and use the sample size $n = 100$.

(a) Display your data in a scatterplot, estimate $\widehat{\beta}_0$ und $\widehat{\beta}_1$ and draw the regression line. Additionally, draw the “true” line using the true parameter values β_0 and β_1.

(b) Take care that your R script is written in a way that you can easily change n and σ^2. Compare the estimated and the “true” regressions lines for different values of n and σ^2. How does the precision of the estimated regression line changes with these values?

(c) Extras:

- How could you ensure that always the same data set is generated?
- Find out how to save your data to a .csv file.