Welcome to R Solution

Tomas Miskov, Valentino Hägg

2022-10-18

This is a short exercise file to get you started with R and R Markdown. If you are not familiar with R, try doing the exercises here before implementing the MM algorithm discussed in the lecture. In our tutorial I will first check with all the R-newcomers if they survived these two exercises, after which we will together go through the implementation of the MM algorithm.

For the R-newcomers: The **welcome_to_R_solution.pdf** is the output you are aiming to create. The **welcome_to_R_solution.Rmd** file shows you exactly how it was created using R Markdown. You are now reading the PDF output of the **welcome_to_R_exercise.Rmd** file. As a good practice, we suggest you open a new R Script and start implementing your solutions to the exercises described in this PDF. Once all your code is running, and the values of your output agree with the **welcome_to_R_solution.pdf**, you can open the **welcome_to_R_exercise.Rmd** and populate the code chunks with your code to replicate the **welcome_to_R_solution.pdf**. Try first only using the PDF file of the solutions for checking your answers, if you really can't solve something, have a look into the Rmd solution file.

Data Manipulation

The following exercises introduce you to the basic commands of R, including the work with vectors, matrices, and some elementary data manipulation. Since importing data is often the first step, let's start with that. Your task in this part is to:

1. Open a fresh R Script file in your chosen folder/location
2. Get your current working directory and print it in the output. Use `getwd()` and `cat()` commands
3. Download the air quality dataset from Canvas, put it in a folder of your choosing, and change the working directory to point to that folder using `setwd()`. If it is in the same folder as your script you do not have to change the working directory. After importing the data in the next step, change back the working directory to where the script is located.
4. Import the air quality data and show the first 5 rows/entries. Use `knitr::kable()` to make a table. Don't forget to include a caption
5. Get the summary statistics for every variable. Use `kable()` again
6. Print the indices of rows that have less than 20mm of rain and airq more than 100. Useful commands can be `as.data.frame()`, `which()`, `cat()`, and `paste()`
7. Finally plot a scatter plot of `airq` vs. `rain`

My current working directory is: C:/Users/misko/OneDrive/Desktop/BDS/Year 2/Supervised Machine Learning Tutorials/bds-ml-ta-materials/MLI/01-r-projects-markdown/tutorial1_illustration

Table 1: The first 5 entries of the Air Quality dataset

airq	vala	rain	coasyes	dens	medi
104	2734.4	12.63	1	1815.86	4397
85	2479.2	47.14	1	804.86	5667
127	4845.0	42.77	1	1907.86	15817
145	19733.8	33.18	0	1876.08	32698
84	4093.6	34.55	1	340.93	6250

Table 2: Summary statistics of the Air Quality dataset

	airq	vala	rain	coasyes	dens	medi
	Min. : 59.0	Min. : 992.9	Min. :12.63	Min. :0.0	Min. : 271.6	Min. : 853
	1st Qu.: 81.0	1st Qu.: 1535.8	1st Qu.:31.02	1st Qu.:0.0	1st Qu.: 365.2	1st Qu.: 3340
	Median :114.0	Median : 2629.8	Median :36.66	Median :1.0	Median : 796.2	Median : 4858
	Mean :104.7	Mean : 4188.5	Mean :36.08	Mean :0.7	Mean : 1728.6	Mean : 9477
	3rd Qu.:126.2	3rd Qu.: 4141.4	3rd Qu.:42.70	3rd Qu.:1.0	3rd Qu.: 1635.2	3rd Qu.: 8715
	Max. :165.0	Max. :19733.8	Max. :68.13	Max. :1.0	Max. :12957.5	Max. :59460

The rows with air quality more than 100 and rain less than 20mm are: 1, 6, 10

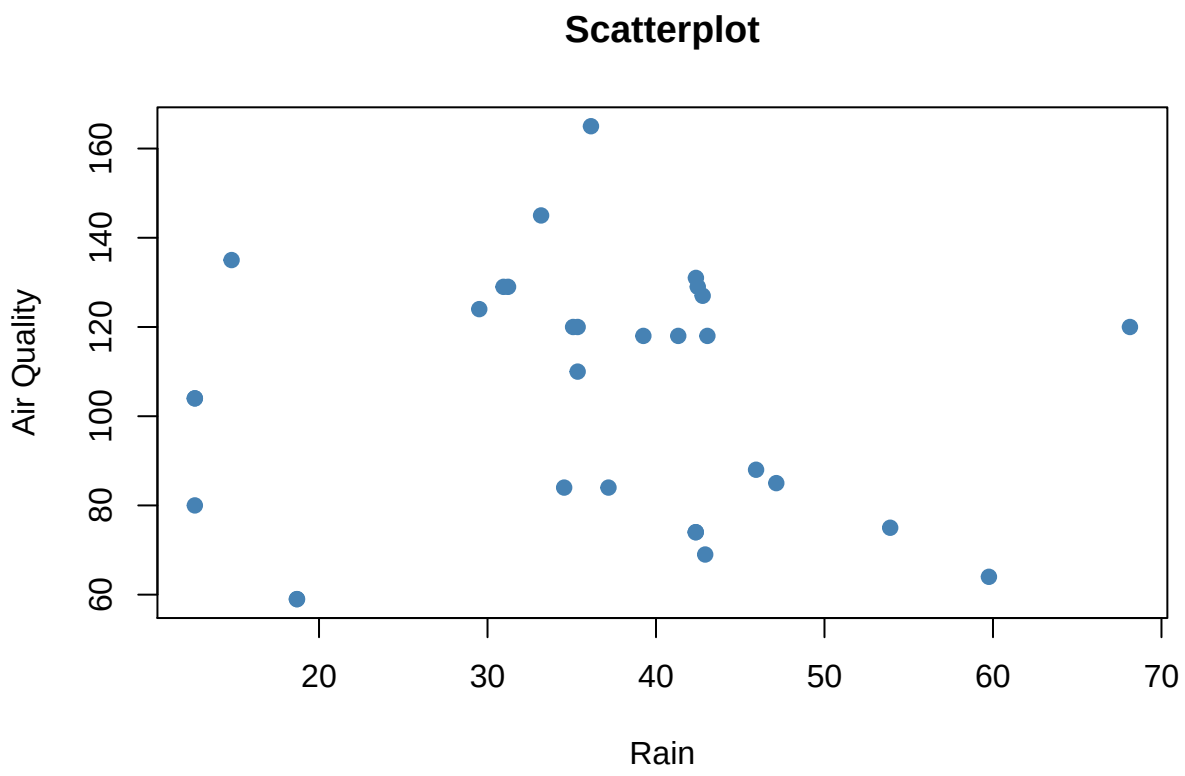

Figure 1: Relationship between air quality and rain

Handling Matrices and Vectors

Hopefully you've struggled a bit with data and now you have the basics of handling a small dataset. Let's now turn to handling vectors and matrices, since that's what R is good at, and what you will do most of the time in this course and beyond. The task is as follows:

1. Set a seed to 123
2. Generate a random matrix $\mathbf{X}$ of 100 observations of 10 explanatory variables. Use normal distribution with mean of 5 and sd of 1.5
3. Generate a random vector of 100 observations of target values $\boldsymbol{y}$, use normal distribution with $\mu = 3$, $\sigma = 2$
4. Normalize the $\mathbf{X}$ values and plot $\mathbf{X}$ vs. $\boldsymbol{y}$
5. Implement an analytical solution to the linear regression in matrix form to predict $\boldsymbol{y}$ from $\mathbf{X}$ (as seen in the lecture slides)
6. Manually calculate R^2 (by calculating TSS and RSS) 7, Compare your vector of coefficients and R^2 with the values obtained from the native `lm()` function

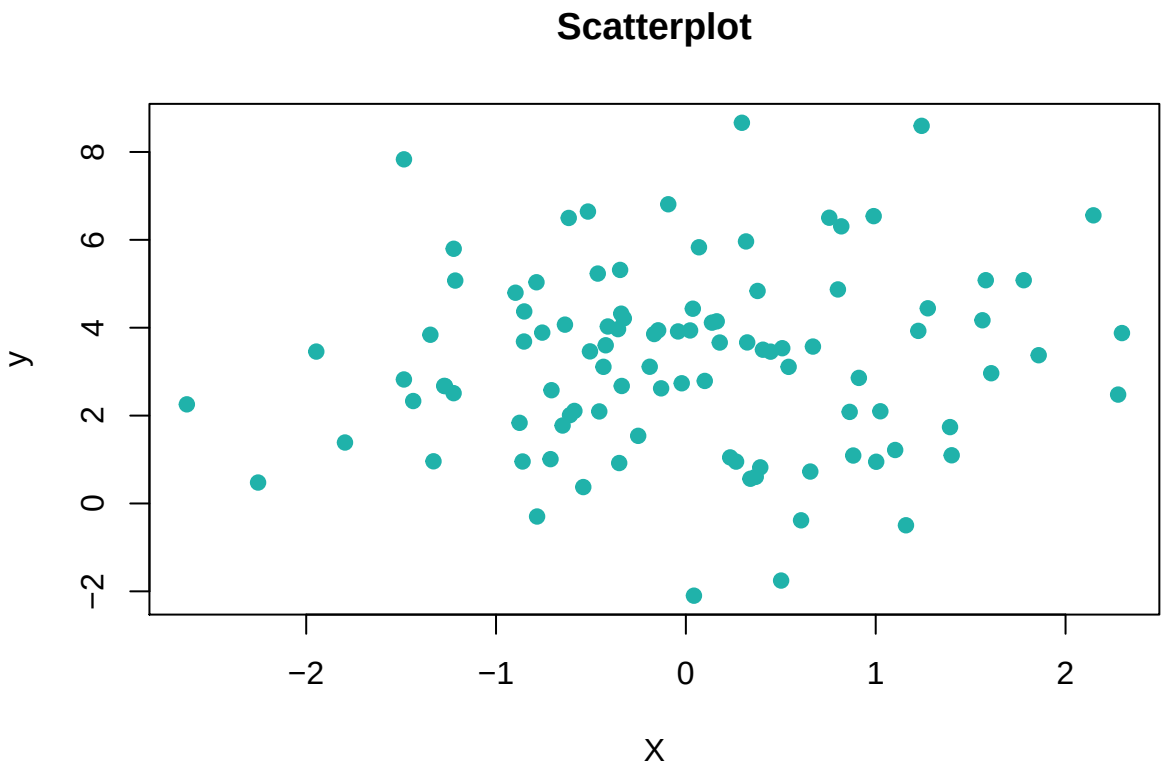

Figure 2: Plot of X vs. y

My beta coefficients have the same values as the coefficients from the lm() function: TRUE

My R^2 is the same as R^2 from lm(): TRUE