Applied Methods in Statistics

Solve Sæbø

2017

Chapter 1

Practical Information

Topics	Exercises
R and R Studio	Exercise 1
Regression Analysis	Exercise 2, Exercise 3
Analysis of Variance	Exercise 4, Exercise 5
Multivariate Statistics	Exercise 6, Exercise 7
Generalized Linear Model	Exercise 8 & 9, Exercise 10
Mixed Effect Models	Exercise 11, Exercise 12

Reference Books

Chapter 2

Getting Started

We will use a dataset containing body measures of about 400 persons as an example set to practice our RStudio skills.

Create New Project

Creating a project allows us to organize the files and related materials duing our study. File => New Project opens a window to create new project. It will be easier to access all the resources, if all the scripts and datasets are within a main folder, i.e. the project folder.

Exercise 1: Data Import

Import from Rdata-file (Workspace file)

Download the data set files "`bodydata.rdata`", "`bodydata.txt`" and "`bodydata.xlsx`" from the fronter pages and save them into your project folder. The files should then be visible under the "files" tab in the bottom-right window in RStudio. A `rdata`-file can be loaded into the workspace either by option 1, 2 or 3 below:

1) Clicking on the filename in the lower right window and answer "yes" to import the data.
2) Clicking the "Open folder icon" under the environment tab in the upper right window. This opens the "Load workspace" dialogue. Locate the data file to load.
3) Load by executing the command (the file must be located in the working directory)

You can change your working directory from main menu "Session > Set working directory" if needed. Any of the above should result in an object called bodydata to appear in the list under "Environment" in the upper right window. This is a data.frame with 407 observations (rows) and 4 variables (columns)

Reading data from `txt` - file

A table of data can be read into an object, here named "`bodydata.from.txt`" from txt-file "`bodydata.txt`" by:

Read the help-file for `read.table` for explanation of extra arguments such as `dec`, `sep` and `header`. Type `?read.table` in console and press return (enter) to access the help or type the name of command, here `data.table` in the search fild of help tab.

Reading data from clipboard ("pasting" copied data into an R object)

You can open the excel-file bodydata.xlsx and copy all variables first, then run:

In Mac, you need to do,

This allows us to get the data from anywhere just after they are copied.

Exercise 2 - Export data to a file

To export an r-object to a file, `write.table` function is used.

Write to txt-file:

Write to memory (copy data)

This will copy the data in r-object on to clipboard which can be pasted anywhere. These data can be directly pasted into an excel-file to create a table.

Save objects as Rdata-file:

Rdata-file allows to save multiple r objects in a file with extension `.rdata`.

This will export bodydata object to a file named `bodydata2.rdata` on your working directory.

Exercise 3 - Exploring the data

Take a look at the top 5 rows of the data

```
  Weight Height Age Circumference
1   65.6  174.0  21          71.5
3   80.7  193.5  28          83.2
4   72.6  186.5  23          77.8
5   78.8  187.2  22          80.0
6   74.8  181.5  21          82.5
```

Take a look at selected rows and column of the data

```
  Weight Age
1   65.6  21
3   80.7  28
4   72.6  23
5   78.8  22
6   74.8  21
```

Here, the `1:5` refers to the number of rows and `c(1, 3)`refers to the number of columns you want from `bodydata`. The output contains the 1 to 5 rows and first and third column for `bodydata`.

Check the dimensions of the data

```
[1] 407   4
```

Here, the first and second item refers to the number of rows and number of columns of bodydata. Similarly, we can use `nrow(bodydata)` and `ncol(bodydata)`to obtain these number individually.

Summary of the data using `summary()` function

```
     Weight          Height           Age         Circumference
 Min.   : 42.00   Min.   :149.5   Min.   :18.00   Min.   : 57.90
 1st Qu.: 58.45   1st Qu.:163.9   1st Qu.:23.00   1st Qu.: 67.95
 Median : 68.60   Median :171.4   Median :27.00   Median : 75.60
 Mean   : 69.19   Mean   :171.3   Mean   :29.91   Mean   : 76.91
 3rd Qu.: 78.80   3rd Qu.:177.8   3rd Qu.:35.00   3rd Qu.: 84.30
 Max.   :108.60   Max.   :198.1   Max.   :67.00   Max.   :113.20
```

The `attach` function

Try to look at the weights by writing Weight and press return. What happens?

Here you will get a ERROR saying that R could not find `Weight` on your workspace.

Attach the data, and try again,

```
  [1]  65.6  80.7  72.6  78.8  74.8  86.4  78.4  62.0  81.6  76.6  83.6
 [12]  90.0  74.6  71.0  79.6  70.0  72.4  85.9  78.8  77.8  81.8  89.6
 [23]  82.8  76.4  63.2  74.8  70.0  72.4  84.1  69.1  67.2  68.6  80.1
 [34]  84.7  73.4  72.1  82.6  88.7  84.1  94.1  74.9  59.1  75.6  86.2
 [45]  75.3  55.2  57.0  61.4  86.8  72.2  71.6  84.8  68.2  66.1  72.0
 [56]  64.6  74.8  70.0 101.6  63.2  78.9  67.7  66.0  68.2  63.9  72.0
 [67]  56.8  74.5  90.9  93.0  80.9  72.7  68.0  72.5  72.5  73.0  70.2
 [78]  70.5 102.3  68.4  75.7  84.5  87.7  86.4  73.2  55.5  58.4  83.2
 [89]  72.7  64.1  72.3  65.0  65.0  88.6  84.1  66.8  75.5  58.0  79.5
[100]  78.6  71.8  72.2  83.6  85.5  90.9  85.9  89.1  75.0  77.7  86.4
[111]  90.9  73.6  76.4  69.1  84.5  69.1 108.6  86.4  80.9  87.7  80.2
[122]  71.4  72.7  76.8  63.6  80.9  85.5  68.6  67.7  66.4 102.3  70.5
[133]  95.9  84.1  71.8  65.9  95.9  91.4  96.8  69.1  82.7  75.5  79.5
[144]  73.6  91.8  85.9  81.8  82.5  80.5  70.0  81.8  84.1  90.5  91.4
[155]  89.1  85.0  69.1  73.6  80.5  82.7  86.4  92.7  93.6  75.0  93.2
[166]  93.2  61.4  83.6  85.5  73.9  66.8  87.3  72.3  88.6 101.4  91.1
[177]  67.3  77.7  76.6  85.0 102.5  77.3  71.8  87.9  94.3  70.9  64.5
[188]  72.3  87.3  80.0  82.3  73.6  74.1  85.9  73.2  76.3  65.9  90.9
[199]  89.1  62.3  82.7  79.1  84.1  83.2  59.0  63.0  59.0  47.6  69.8
[210]  75.2  55.2  62.5  42.0  50.0  49.2  73.2  68.8  50.6  57.2  87.8
[221]  72.8  54.5  59.8  67.3  47.0  46.2  55.0  83.0  54.4  45.8  53.6
[232]  52.1  67.9  56.6  62.3  58.5  54.5  50.2  60.3  58.3  56.2  50.2
[243]  72.9  59.8  61.0  69.1  55.9  46.5  60.0  60.3  52.7  74.3  62.0
[254]  73.1  80.0  54.7  75.7  61.1  55.7  48.7  52.3  50.0  59.3  62.5
[265]  55.7  54.8  45.9  69.4  64.8  71.6  52.8  59.8  49.0  50.0  69.2
[276]  63.4  58.2  45.7  52.2  48.6  55.6  66.8  59.4  53.6  69.0  58.4
[287]  56.2  70.6  59.8  72.0  65.2  56.6 105.2  51.8  63.4  59.0  47.6
[298]  55.2  45.0  54.0  50.2  44.8  58.8  56.4  67.2  53.8  54.4  58.0
[309]  54.8  43.2  46.4  64.4  62.2  55.5  57.8  54.6  59.2  52.7  53.2
[320]  64.5  51.8  56.0  63.2  59.5  56.8  50.0  72.3  55.0  60.4  69.1
[331]  84.5  55.9  69.5  76.4  58.6  66.8  56.6  58.6  55.9  56.8  60.0
[342]  58.2  72.7  54.1  49.1  75.9  55.0  55.0  65.5  65.5  58.6  55.2
[353]  62.7  56.6  53.9  63.2  73.6  62.0  63.6  53.2  53.4  55.0  70.5
[364]  54.5  55.9  59.0  47.3  67.7  80.9  70.5  60.9  63.6  54.5  59.1
[375]  52.7  62.7  66.4  67.3  63.0  73.6  62.3  57.7  55.4 104.1  77.3
[386]  64.5  61.4  58.2  63.6  53.4  54.5  53.6  60.0  73.6  61.4  55.5
[397]  60.9  60.0  46.8  64.1  63.6  67.3  68.2  61.4  76.8  71.8  67.3
```

Attach makes the variables "visible" from outside the data frame. The opposite function is `detach()`. After attaching a data frame, R can access its variables just like another R object.

Exercise 4 - Subsets of data and logical operators

We can perform logical test and get `TRUE` or `FALSE` as result. Make a variable called `isHeavy` by,

Take a look at this variable, what is it?

```
[1] FALSE  TRUE FALSE FALSE FALSE  TRUE
```

Yes, it is a vector of `TRUE` and `FALSE` with same length as `Weight`. Here the condition has compared each element of `Weight` results `TRUE` if it is greater than 80 and `FALSE` if it is less than 80.

Identify the elements

We can identify which observations that are heavy by the `which()` function

How many are heavy?

```r
length(HeavyId)
```

```
[1] 94
```

In the similar manner, identify who are taller than 180 and save this as an object called `isTall`.

```r
isTall <- Height > 180
```

Here, you can use `length` function as above to find how many person are taller than 180.

Try

How is this computation done?

Here `isHeavy` and `isTall` contains `TRUE` and `FALSE`. The multiplication of logical operator results a logical vector with `TRUE` only if both the vectors are `TRUE` else `FALSE`.

Alternatively

The `&` operator result `TRUE` if both `isHeavy` and `isTall` are `TRUE` else, `FALSE` which is same as previous.

Subsetting data frame

Create a subset of the data called `bodydataTallAndHeavy` containing only the observations for tall and heavy persons as defined by isBoth.

```
bodydataTallAndHeavy <- bodydata[isBoth, ]
```

For other logical tests see help file `?Comparison`

Create a random subset of 50 observations by

1) sampling a vector of random observation numbers using the `sample()` function, and then

Here, 50 rows are sampled from the total number of rows and the index of the selected rows are saved on vector `idx`.

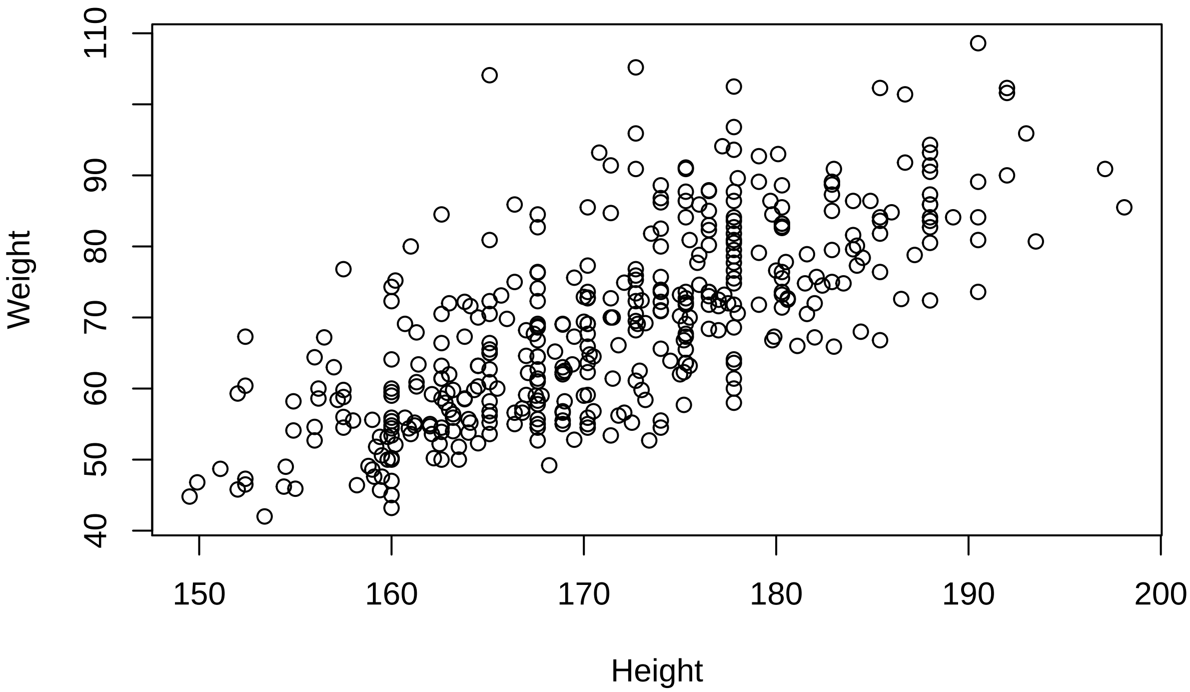
2) using the vector to select the rows to create a new data set called `bodydataRandom`.

```
bodydataRandom <- bodydata[idx, ]
```

Exercise 5 - Graphics

Plot the heights versus the weights for all observations in bodydata.

```
plot(x = Height, y = Weight)
```



Spice up the plot

Check out the presentation for lesson 1 to see how to spice up the plot

Explore the `?par` help file

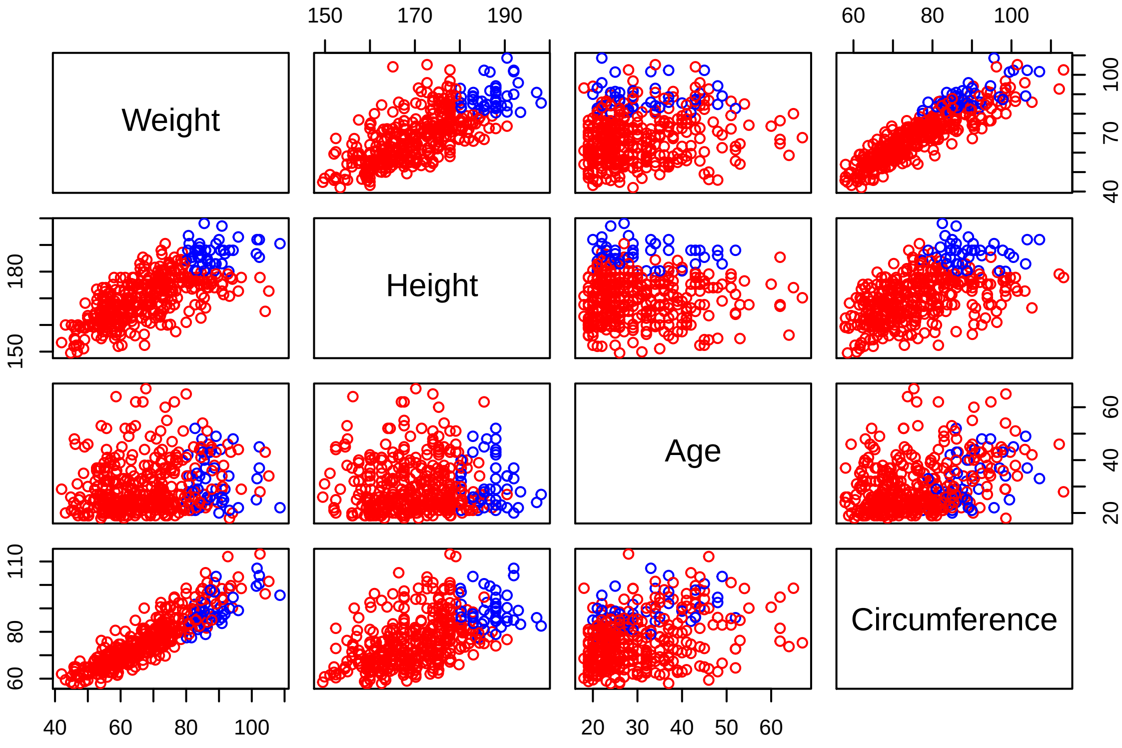
Use the `isBoth` variable to create a color vector

Here, `isHeady & isTall` returns a logical vector. The `ifelse` function returns `blue` if `TRUE`and `red` if `FALSE` for each elements of the logical vector. The colors are then used in the plot so that all the Heavy and Tall person will be colored "blue" and rest as "red".

Use "`mycolors`" in the `col` argument of the plot function to mark the tall and heavy individuals

Plot all variables against each other

```
pairs(bodydata, col = mycolors)
```



Which variables seem to be most correlated to each other?

Here, `Weight` and `Circumference` seems to have highest correlation.

Which variables are least correlated to each other?

`Age` and `Height` variables seems to have least correlation.

Check by,

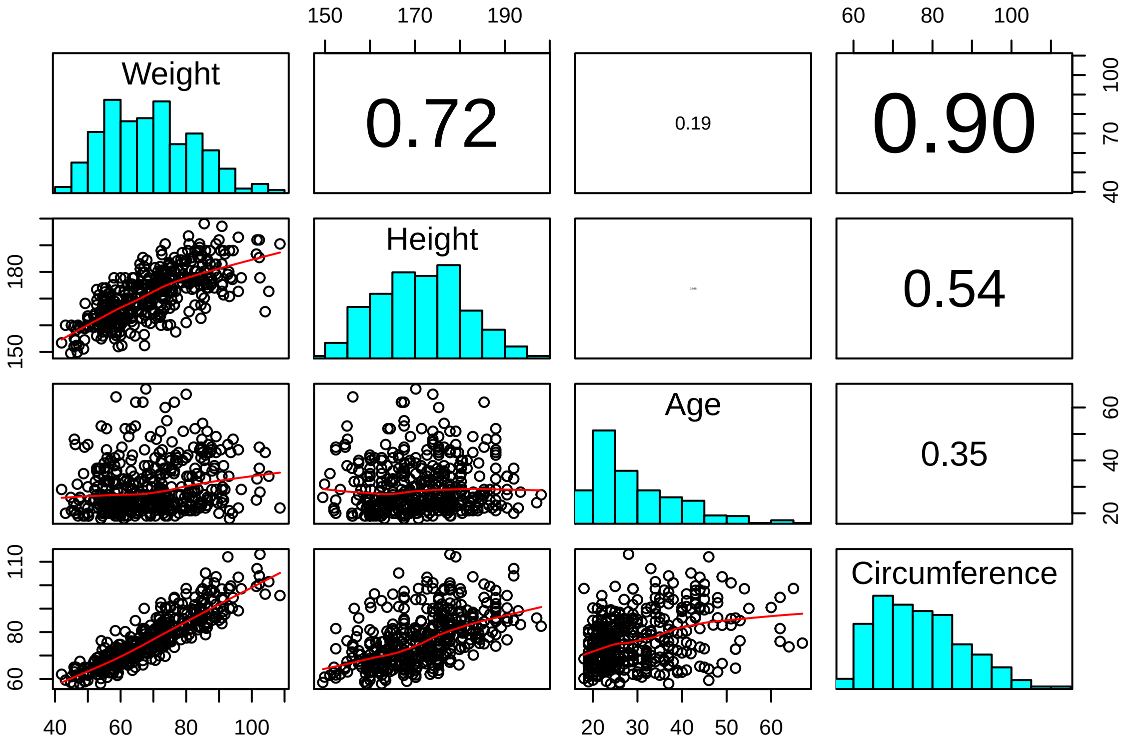
```
                 Weight     Height        Age Circumference
Weight        1.0000000 0.72080269 0.18701340     0.8994638
Height        0.7208027 1.00000000 0.04822479     0.5447980
Age           0.1870134 0.04822479 1.00000000     0.3547390
Circumference 0.8994638 0.54479800 0.35473898     1.0000000
```

This returns the correlation matrix for the variables, and the guess made earlier true. Further, check out the help file for pairs, and the examples at the end. Try to make a pairs plot with scatter plots with smoothed lines in the lower left triangle, histograms on the diagonal, and correlation numbers in the upper right triangle.

Lets first create a function which create histogram. The function will later be used in the `pairs` function to create its diagonal plots.

Now, create a function that will display correlation on `pairs` plot.

Now, the above functions are implemented on the `pairs` plot,



Here the `panel.smooth` deals with the smooth line on the lower panel of `pairs` plot.

Note:: Chapter 5 of the R book contains numerous examples of graphics.

Chapter 3

Group Exercises

Exercise One

In the file Audi.Rdata in the Data folder on Fronter you find sales prices and technical data on 30 cars of type Audi A4. The data were collected on Feb 15th 2017. The variables are:

Variable		Description
`Price`	:	Price of the car (In 1000 NOK)
`Km`	:	Distance driven (in 1000 Km)
`Hk`	:	Horse power
`Transition`	:	Transition system (categorical). M=manu
`Volume`	:	Cylinder volume
`Fuel`	:	Fuel type (categorical). D=Diesel, G=Gas
`CO2`	:	CO2-emission (g/km)
`Weight`	:	The weight of the car
`Year`	:	Production year
`Age`	:	Years since production (=2017 – year)

Consider Price as the response variable and all other variables (but not year) as candidate predictor variables.

a) Fit the full model including all candidate predictors. Write up the estimated model. Discuss the effect estimates. How should they be interpreted?

The full model with all candidate predictors with price as response can be written as,

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ &\quad + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \epsilon, \\ \text{where, } &\epsilon \sim \text{NID}(0, \sigma^2) \end{aligned}$$

Here,

...

The least square estimate for this model can be obtained from R as,

```
model.1 <- lm(Price ~ Km + Hk + Transition + Volume + Fuel + CO2 + Weight + Age,
              data = Audi)
summary(model.1)
```

```
Call:
lm(formula = Price ~ Km + Hk + Transition + Volume + Fuel + CO2 +
    Weight + Age, data = Audi)

Residuals:
    Min      1Q  Median      3Q     Max
-48.372 -14.146   4.172  13.381  55.476

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   394.9427   378.4255   1.044  0.30851
Km             -0.4645     0.1551  -2.995  0.00690 **
Hk              1.0403     0.5036   2.066  0.05139 .
Transition(A)   3.6993     8.4920   0.436  0.66756
Volume         66.6204   103.1563   0.646  0.52539
Fuel(D)       -11.3385    14.8522  -0.763  0.45370
CO2            -0.1983     0.6347  -0.312  0.75784
Weight         -0.1848     0.2654  -0.696  0.49380
Age           -15.5632     4.6665  -3.335  0.00314 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 30.97 on 21 degrees of freedom
Multiple R-squared: 0.8964,
Adjusted R-squared: 0.857
F-statistic: 22.72 on 8 and 21 DF,  p-value: 0.00000001097
```

The estimated model can be written as,

$$
\begin{aligned}
\hat{y} = (394.94) + (-0.46)x_1 + (1.04)x_2 + (3.7)x_3 + (66.62)x_4 \\
+ (-11.34)x_5 + (-0.2)x_6 + (-0.18)x_7 + (-15.56)x_8
\end{aligned}
$$

b) Use variable selection methods or the best subsets approach to identify a good reduced model. Discuss the results.

c) Check the model assumptions of your final model.

d) Are there any influential observations?

e) Redo exercises a) to d), but start with a full model including `Age`, `Hk` and `Km` and all second order interactions between these variables. In R Commander you may specify the model like this:

```
Price ~ (Age + Hk + Km)^2
```

(*This notation means: Include all predictors up to second order interactions.*) Discuss the final model estimate.

f) If there is extra time: Can you find an even better model than the one you found in e?

Bibliography