Dr. Marcel Dettling Applied Statistical Regression AS 2011

Series 1

1. The goal of this exercise is to get acquainted with different abilities of the `R` statistical software. It is recommended to use the distributed `R` tutorial as a guide.

`R` contains more than 50 datasets and more can be loaded using optional packages. The package `VR` is depending on the package `MASS` which contains the dataset `survey`. This dataset comprises of measurements and answers taken from 237 students of statistics at the university of Adelaide. The following variables are available

Sex	gender of student
Wr.Hnd	span width in cm (from thumb to pinky) of the writing hand
NW.Hnd	span width in cm (from thumb to pinky) of the non-writing hand
W.Hnd	writing hand
Fold	When folding your arms - which one is on top?
Pulse	beats per minute
Clap	When clapping your hands - which on is on top?
Exer	How often do you exercise?
Smoke	How often do you smoke?
Height	body length in cm
M.I	Preference of either metric (cm/m) or imperial (feet/inches) units?
Age	age in years

> **library(MASS)**	makes the datasets of the `MASS` package available **PC: Install first the package** `VR`
> **data()**	shows a list of all available datasets
> **help(survey)**	gives a description of the dataset `survey`
> **data(survey)**	makes the dataset `survey` available

Useful functions to get a first overview of the dataset:
str(survey), **summary(survey)**, **table(survey$Sex)**, **table(survey$Sex, survey$Smoke)**
The notation `survey$Smoke` accesses the variable `Smoke` in the dataset `survey`.

> **attach(survey)**	puts the dataset `survey` on level 2 of the list of available objects. The working directory is on level 1. The variables in the dataset `survey` can now be accessed directly with their names, i.e. instead of typing `survey$Smoke` you may access the variable directly with `Smoke`.

Dealing with missing values (`NA`):

> **mean(Pulse)**	result is `NA`
> **mean(Pulse, na.rm=T)**	the missing values are removed from the calculation of the mean
> **na.omit(Pulse)**	all missing values are removed
> **Pulse[!is.na(Pulse)]**	same as above, but generated *by hand*

Useful functions for graphics:

> **hist(Height)**	histogram
> **boxplot(Height)**	boxplot
> **boxplot(split(Height, Sex))**	boxplots of two variables
> **boxplot(Height[Sex=="Female"],Height[Sex=="Male"])**	boxplots
> **plot(Wr.Hnd,NW.Hnd)**	scatter plot
> **plot(Sex,Height)**	?

> **detach(survey)** disconnects the dataset `survey` from level 2, i.e. variables can no longer by accessed directly, but only using `$` or `[·,·]`:
> **plot(survey$Wr.Hnd,survey$NW.Hnd)** or **plot(survey[,2],survey[,3])**.

Selecting observations, i.e. only the first 50:
> **plot(survey[1:50,2],survey[1:50,3])**

Do not forget about the online help:
> **help(survey)**
> **help(plot)**
...

Now analyse the dataset `survey` using descriptive methods. Therefore produce tables and contingency tables of the categorical variables and calculate location and deviation properties for the continuous variables. Provide suitable graphical representations. Comment on the distributions. Are there any outliers?

Answer the following questions:

a) Is the span width of the writing hand in general larger than the span width of the non-writing hand?

b) Do the two oldest students smoke?

c) Which factors might have an influence on the student's pulse?

d) It is generally believed that the pulse of an individual decreases with increasing age. The function `lm` fits a linear regression. Investigate the output of the following code:

> **Agejung <- Age[Age<30], Pulsejung <- Pulse[Age<30], plot(Agejung,Pulsejung)**

Comment on the output. What does the above code do?

> **lmobj <- lm(Pulsejung ~ Agejung),plot(Agejung,Pulsejung),abline(lmobj)**

Preliminary discussion: Monday, October 03.

Deadline: —.