Universitetet i Bergen Summer 2019
STAT200 - Applied Statistics

Hand-in assignment

This hand-in assignment deals with linear models which you have already worked with in the previous assignments. Since it needs to be handed in, please read the following instructions carefully:

1. After downloading the data, store all data sets in a folder named `Data`. This folder has to be located in the same directory on your computer as the file containing your `R`-scripts.

2. Create only one file for all exercises, which is of the following structure: for each exercise (and sub-exercise), first the `R`-code you programmed for solving the respective (sub-)exercise, followed by your replies in comment form (use # for not producing errors). Example:

```
# Exercise 1
# ----------

# 1.1
# ---

x <- rnorm(1000)
hist(x)
shapiro.test(x)
# I created a Gaussian sample with 1000 obvservations, plotted a histogram, and
# carried out the Shapiro-Wilk test for normality.  The resulting p-value...

# 1.2
# ---
...
```

3. When finished, submit your result via MittUiB/Canvas, a task will be created for this purpose. You are allowed to upload one file, which has to be of format `yourlastname.yourfirstname.r`. Pay attention, only files ending on `.r` are accepted by the system.

4. Execution of the script must not produce any errors. Parts resulting in errors will be completely ignored and graded with zero points.

5. Deadline for submission is 25.3.2019 (Monday) at 23h00. You will receive your result (pass / fail) approximately one week later.

6. Any questions requesting further information on how to solve / interpret the exercises will be ignored by me and the assistants. Only if you are certain you have discovered a mistake, please contact us.

7. As usually, we will run the hand-ins through standard plagiarism checks. If such cases are identified, all affected hand-ins will be graded as fail.

Exercise 1

The data set gsa.csv contains information on consumer behavior during their holidays. It consists of the following variables:

- income - the income of the person renting an accommodation.
- country - the country where the rental took place.
- expenditure - the amount spent per week.
- accomodation - the type of accommodation.
- year - the year during which the rental took place.

Your tasks are the following.

1. A first working hypothesis is that the expenditures as response variable are somehow affected by the predictors 'income' and 'country', further details are not known. Carry out a full analysis of the data set, using techniques that you have learned in the lecture and previous assignments.
2. As a second step, are you able to include the remaining variables in a reasonable way? How does this affect the results?
3. Assume, hypothetically, that you are given the additional information that the 1100 data points collected belong to 110 households, each of which was observed over 10 successive years. Would you carry out the same analysis, or do something else (if so, what and why)? This question is purely theoretical.

Exercise 2

The data set 'ELISA red stats.xlsx' reports protective antibody levels of a species of cleaner fish, lumpsuckers (Cyclopterus lumpus), which were followed after vaccination with two different vaccines. Fish were sampled every 100 degree days (d.d.) for a period of 600 d.d. (this means 60 days at 10 degrees Celsius). Blood samples were taken from 40 fish for each vaccine at each time point, and blood analysis carried out via ELISA (enzyme-linked immunosorbent assay). Essentially, the more efficient a vaccine is, the higher the levels of protective antibodies that are generated, until a saturation point is reached. The data file contains the following variables:

- Vaccine - the type of vaccine. Past 1 corresponds to the first vaccine (non-concentrated vaccine), and Past 2 to an alternative vaccine (concentrated vaccine).
- Day - corresponds to 100 until 600 d.d. from vaccination.
- Abs - the absorbency value resulting from the ELISA for each sample analyzed.

Your tasks are the following.

1. Read in the data and represent them visually. In case you fail to read in an Excel file, save it e.g. as csv-type file. Note that the colleague who provided the data has marked some potentially outlying values due to experimental errors. Treat these outliers within R, do not modify the original data file.

2. Carry out an analysis of the data set. There is not real working hypothesis, except for that the two vaccines may somehow be different. A starting point would be a simple two-way ANOVA.

3. Try to also investigate other models. Is it possible (reasonable) to treat the d.d. as numerical value? Or does the introduction of dummy variables for determining certain time periods make sense. You may develop your own hypothesis based on the visual representation of the data.