EDDA: Assignment 2

Throughout this assignment tests should be performed using a level of 0.05, unless otherwise specified.

Exercise 1. Moldy bread

If left alone bread will become moldy, rot or decay otherwise. To investigate the influence of temperature and humidity on this process, the time to decay was measured for 18 slices of white bread, which were placed in 3 different environments and humidified or not. The data are given in the file `bread.txt`, with the first column time to decay in hours, the second column the environment (cold, warm or intermediate temperature) and the third column the humidity.

a) The 18 slices came from a single loaf, but were randomized to the 6 combinations of conditions. Present an R-code for this randomization process.

b) Make two boxplots of `hours` versus the two factors and two interaction plots (keeping the two factors fixed in turn).

c) Perform an analysis of variance to test for effect of the factors `temperature`, `humidity`, and their interaction. Describe the interaction effect in words.

d) Which of the two factors has the greatest (numerical) influence on the decay? Is this a good question?

e) Check the model assumptions by using relevant diagnostic tools. Are there any outliers?

Exercise 2. Search engine

A researcher is interested in the time it takes a student to find a certain product on the internet using a search engine. There are three different types of interfaces with the search engine and especially the effect of these interfaces is of importance. There are five different types of students, indicating their level of computer skill (the lower the value of this indicator, the better the computer skill of the corresponding student). Fifteen students are selected; three from each group with a certain level of computer skill. The data is given in the file `search.txt`. Assume that the experiment was run according to a randomized block design which you make in a). (Beware that the levels of the factors are coded by numbers.)

a) Number the selected students 1 to 15 and show how (by using `R`) the students could be randomized to the interfaces in a randomized block design.

b) Test the null hypothesis that the search time is the same for all interfaces. What type of interface does require the longest search time? For which combination of skill level and type of interface is the search time the shortest? Estimate the time it takes a typical user of skill level 3 to find the product on the website if the website uses interface 3.

c) Check the model assumptions by using relevant diagnostic tools.

d) Perform the Friedman test to test whether there is an effect of interface.

e) Test the null hypothesis that the search time is the same for all interfaces by a one-wayANOVA test, ignoring the variable `skill`. Is it right/wrong or useful/not useful to perform this test on this dataset?

Exercise 3. Feedingstuffs for cows

In a study on the effect of feedingstuffs on lactation, a sample of nine cows were fed food of two types, and their milk production was measured. All cows were fed both types of food, during two periods, with a neutral period in-between to try and wash out carry-over effects. The order of the types of food was randomized over the cows. The observed data can be found in the file `cow.txt`, where `A` and `B` refer to the types of feedingstuffs.

a) Test whether the type of feedingstuffs influences milk production using an ordinary "fixed effects" model, fitted with `lm`. Estimate the difference in milk production.

b) Repeat a) by performing a mixed effects analysis, modelling the cow effect as a random effect (use the function `lmer`). Compare your results to the results found by using a fixed effects model. (You will need to install the `R`-package `lme4`, which is not included in the standard distribution of `R`.)

c) Study the commands:

```
> attach(cow)
> t.test(milk[treatment=="A"],milk[treatment=="B"],paired=TRUE)
```

Does this produce a valid test for a difference in milk production? Is its conclusion compatible with the one obtained in a)? Why?

Exercise 4. Jane Austen

Stochastic models for word counts are used in quantitative studies on literary styles. Statistical analysis of the counts can, for example, be used to solve controversies about true authorships. Another example is the analysis of word frequencies in relation to Jane Austen's novel *Sanditon*. At the time Austen died, this novel was only partly completed. Austen, however, had made a summary for the remaining part. An admirer of Austen's work finished the novel, imitating Austen's style as much as possible. The file `austen.txt` contains counts of different words in some of Austen's novels: chapters 1 and 3 of *Sense and Sensibility* (stored in the `Sense` column), chapters 1, 2 and 3 of *Emma* (column `Emma`), chapters 1 and 6 of *Sanditon* (both written by Austen herself, column `Sand1`) and chapters 12 and 24 of *Sanditon* (both written by the admirer, `Sand2`).

a) Discuss whether a contingency table test for independence or for homogeneity is most appropriate here.
b) Using the given data set, investigate whether Austen herself was consistent in her different novels. Where are the main inconsistencies?
c) Was the admirer successful in imitating Austen's style? Perform a test including all data. If he was not successful, where are the differences?

Exercise 5. Expenditure on criminal activities

The data in `expensescrime.txt` were obtained to determine factors related to state expenditures on criminal activities (courts, police, etc.) The variables are: `state` (indicating the state in the USA), `expend` (state expenditures on criminal activities in $1000), `bad` (crime rate per 100000), `crime` (number of persons under criminal supervision), `lawyers` (number of lawyers in the state), `employ` (number of persons employed in the state) and `pop` (population of the state in 1000). In the regression analysis, take `expend` as response variable and `bad`, `crime`, `lawyers`, `employ` and `pop` as explanatory variables.

a) Make some graphical summaries of the data. Investigate the problem of potential and influence points, and the problem of collinearity.
b) Fit a linear regression model to the data. Use both the step-up and the step-down method to find the best model. If step-up and step-down yield two different models, choose one and motivate your choice.
c) Check the model assumptions (of the resulting model from b)) by using relevant diagnostic tools.