SDA 2019 — Assignment 11

Categorical data analysis – hand in Exercises 11.1 and 11.2 until Tuesday 7th of May, 11.59pm.

Linear regression – use Exercises 11.3 and 11.4 to prepare the Exercises of Assignment 12; in Assignment 12 you will further analyze both considered datasets. Do not hand in your solution to Exercises 11.3 and 11.4. Use your time wisely!

For Exercises 11.1 and 11.2 you may use the *R*-function `chisq.test` and the functions `bootstrapcat` and `maxcontributionscat`.[1] When performing statistical tests, *clearly* state the null hypothesis and test statistic together with its distribution under the null hypothesis.

Make a concise report of *all* your answers in *one single PDF file*, with only *relevant R* code *in an appendix*. It is important to make clear in your answers how you have solved the questions. Graphs should look neat (label the axes, give titles, use correct dimensions etc.). Multiple graphs can be put into one figure using the command `par(mfrow=c(k,l))`, see `help(par)`. Sometimes there might be additional information on what exactly has to be handed in. **Read the file AssignmentFormat.pdf on Canvas carefully**.

Exercise 11.1 The file `nausea.txt` contains data about post-operative nausea after medication against nausea. The patients, who complained about post-operative nausea, were randomly assigned to one of the different medicines or to a placebo. One of the medicines, Pentobarbital, was administered in two different doses. The first column in the file contains the **total** number of patients that were given the medicine or placebo, and the second column contains the number of cases of nausea (after medication) in that group. For all tests considered below, we choose the significance level $\alpha = 5\%$.

a. Compute for each medicine (and the placebo) the number of patients that did not show nausea. Investigate for each of the four medicines whether it decreases post-operative nausea when compared to the placebo using Fisher's exact test.

In parts b–g consider the whole data set, and use chi-square tests.

b. Which of the three models II A, II B and II C is most suitable for these data?

c. Investigate in a suitable way whether there is a dependency between the variables 'treatment' and 'incidence of nausea'.
Hint: make sure that the null hypothesis, that you are testing, is in line with what can be tested within the model you specified in b.

d. Compute the contributions and the standardized residuals for the test(s) that you performed in part c. Which 'categories' stand out?

e. Compare the medicines (in pairwise comparisons) to the placebo. Do the results agree with what you found in part c?

Hand in: your answers to all parts.

[1] You can find these functions in the file `functions_Ch7.txt`.

Exercise 11.2 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*, chapters 1, 2 and 3 of *Emma*, chapters 1 and 6 of *Sanditon* (both written by Austen herself, Sand1) and chapters 12 and 24 of *Sanditon* (both written by the admirer, Sand2). For all tests considered below, we choose the nominal level $\alpha = 10\%$.

a. Which of the three models, II A, II B and II C is most suited for these data?

b. Investigate using these data whether Austen herself was consistent in her different novels. In case you find that Austen was not consistent, find out where the main inconsistencies are.

c. Was the admirer successful in imitating Austen's style? If not, where are the differences?

d. Check the obtained p-value from the tests that you used in parts b and c with the bootstrap method, using the R-function `chisq.test`. What is your conclusion?

For bootstrap procedures for statistics other than the standard chi-square statistic you can use the function `bootstrapcat`. Such a procedure can be rather time consuming. Adapt the value of B such that it is feasible. For all tests considered below, use the two-sided alternative hypotheses. Also, use the same nominal level α as above.

e. Test the same null hypothesis as in part c, but now with a bootstrap procedure using the statistic $T =$*'The largest of the absolute values of the contributions'*. To compute the observed value of this statistic the R-function `maxcontributionscat` available on Canvas can be used.

f. Do the same as in part e with the test statistic $T =$ *'The fourth largest of the absolute values of the contributions'*. Perform the bootstrap test. To compute the observed value of these statistics, one could, for instance, adjust the function `maxcontributionscat`.

g. Do the results of parts e and f agree with those of part c?

Hand in: your answers to all parts.

For these exercises the *R* function `lm` is needed to fit linear models. The data to be analyzed should be in a `data.frame` format, see the first exercise.[2]

Exercise 11.3 Aerial survey methods are used to estimate the number of snow geese in their summer range areas west of Hudson's Bay in Canada. To obtain the estimates, small aircrafts fly over the range and, when a flock of snow geese is spotted, an experienced observer estimates the number of geese in the flock. To investigate the reliability of this method, an experiment was conducted. An airplane carrying two observers flew over 45 flocks, and each observer independently estimated the number of geese in the flock. Also, a photograph of the flock was taken so that an exact count of the number in the flock could be obtained. The data are contained in the file `geese.txt`.

a. Draw scatter plots of the observer counts (Y) versus the photo count (x). Do these graphs suggest a simple linear regression model might be appropriate?

b. Perform the linear regression for the two observers separately. Fit the parameters and test the hypothesis: $\beta_1 = 0$ against the alternative: $\beta_1 \neq 0$ with significance level 0.05 in each model.

Do not hand in your results. This exercise serves as a preparation for a more thorough data analysis of this dataset in Assignment 12.

Exercise 11.4 This exercise concerns data measured by the Los Angelos Pollution Control District. This agency attempts to construct statistical models to predict pollution levels. The file `airpollution.txt` contains the maximum level of an oxidant (a photochemical pollutant) and the morning averages of four meteorological variables: wind speed, temperature, humidity and insolation (a measure for the amount of sunlight). The data cover 30 days during one summer. Investigate which explanatory variables need to be included into a linear regression model with `oxidant` as the response variable by performing the steps below.

a. Make scatter plots of the four candidate explanatory variables against each other and against the response variable (see the *R*-function `pairs()`). Interpret the plots. Do you judge a linear model to be useful here?

b. Determine for each of the explanatory variables the simple linear regression model.

Do not hand in your results. This exercise serves as a preparation for a more thorough data analysis of this dataset in Assignment 12.

[2]You can find this and other functions that we will use next time in the file `functions_Ch8.txt`.