COURSEWORK ASSIGNMENT A - 2018

CS4125 – SEMINAR RESEARCH METHODOLOGY FOR DATA SCIENCE

The result of this coursework assignment should be combined into a single PDF report. The R code written for the coursework should be included as an appendix, as well as the output of analyses.

Part 1 – Design and set-up of true experiment

Write a plan for conducting an experiment on group of human test subjects. As a group you are allowed to select your own topic for this experiment. The plan should include the following items.

- The motivation for the planned research. (Max 250 words)
- The theory underlying the research. (Max 250 words) Preferable based on theories reported in literature
- Research questions that will be examined in the experiment (or alternatively the hypothesis that will be tested in the experiment)
- The related conceptual model, this model should include:
    - Independent variable(s)
    - Dependent variable
    - Mediating variable (at least 1)
    - Moderating variable (at least 1)
- Experimental Design (the study should have a true experimental design)
- Experimental procedure (how the experiment will be executed step by step)
- Measures
- Participants
- Suggested statistical analyses

Part 2 – Generalized linear models

Question 1 Twitter sentiment analysis (Between groups – single factor)
Analyzing Twitter tweets about a specific topic or person, it is possible to a get an overall sense the sentiment of these tweets. This is done by counting the number of positive and negative words in a tweet. The main aim of this question is that you compare the sentiment of the tweets related to at least 3 famous individuals (i.e. celebrities) that are often the topic of discussion on Twitter (in English). The file "TwitterAnalysis.R" shows how you can obtain tweets automatically. This program uses the following file which you need to place in you working directory: sentiment3.R, negative-words.txt, and positive-words.txt.

For the analysis you need to have a twitter account to create a so called "twitter app" on apps.twitter.com. Once you have done this, obtain information under

"Keys and Access Tokens" and enter these in your own file with your personal twitter variables. For this you can use the template file "your_twitter.R".

Once you have done this, conduct the following analyses on the obtained data set.

1. Make a conceptual model for the following research question: Is there a difference in the sentiment of the tweets related to the different celebrities?
2. Analyze the homogeneity of variance of sentiments of the tweets of the different celebrities
3. Graphically examine the variation in tweets' sentiments for each celebrity (e.g. histogram, density plot etc.)
4. Graphically examine the mean sentiments of tweets for each celebrity
5. Use a linear model to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets.
6. If a model that includes the celebrity is better in explaining the sentiments of tweets than a model without such predictor, conduct a post-hoc analysis with e.g. Bonferroni correction, to examine which of celebrity tweets differ from the other celebrity tweets
7. Write a small section for a scientific publication, in which you report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.
8. Include the annotated R script (excluding your personal Keys and Access Tokens information) in the appendix of the report.

Question 2 – Website visits (between groups – Two factors)

For this question you have to use the data file webvisit[x].csv. There are 3 versions of this data set (0,1, and 2). To determine the version your group has to select, add up the age (in years, at the first official day of the course) of the group members and take modulo 3 of this number. The obtained number is the version your group has to complete.

The file represents data obtained from a webserver from a company X. The company runs an A-B study to test two versions of their website (0 = old, 1 = new version. The company targets two markets and therefore has two web portal entries (0=consumers, 1 = companies). For each visit to their website, the data file shows the number of pages the visitor visited. The aim of the analysis is to examine whether the version of the website, the portal, or combination of the two had an impact on number of pages visited.

1. Make a conceptual model underlying this research question
2. Graphically examine the variation in page visits for different factors levels (e.g. histogram, density plot etc.)
3. Statistically test if variable page visits deviates from normal distribution
4. Conduct a model analysis, to examine the added values of adding 2 factors and interaction between the factors in the model to predict page visits.

5. If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail.
6. Write a small section for a scientific publication, in which you report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.
7. Include annotated R script in the appendix of the report.

Question 3: Linear regression analysis

Select a data set from the following sites and conduct a linear regression:

- (http://www.stat.ufl.edu/~winner/datasets.html)
- http://support.minitab.com/en-us/datasets/ (install trail version of minitab to open the minitab file. Copy data to excel file and open this in R)
- Dutch CBS Statistics Netherlands http://statline.cbs.nl/Statweb/?LA=en

The data you select should meet the following requirements:

- n > 100
- at least 3 independent variables of interval (or ratio) level
- a dependent variable of interval (or ratio) level and which is reasonable normally distributed
- Independence of observations

Conduct the following analysis on the data set:

1. Make a conceptual model underlying this research question
2. Graphical analysis of the distribution of the dependent variable, e.g. histogram, density plot
3. Scatter plots between dependent variable and the predictor variables
4. Conduct a multiple linear regression (including confidence intervals, and beta-values)
5. Examine assumptions underlying linear regression. E.g collinearity and analyses of the residuals, e.g. normal distributed (QQ plot), linearity assumption, homogeneity of variance assumption. Where possible support examination with visual inspection.
6. Examine effect of single cases on the predicted values (e.g. DFBeta, Cook's distance)
7. Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.
8. Include annotated R script in the appendix of the report.

Question 4 Logistic regression analysis

Select a data set from the following sites and conduct a logistic regression:

- http://www.stat.ufl.edu/~winner/datasets.html
- http://support.minitab.com/en-us/datasets/ (use foreign and read.mtp function to read minitab files)
- Dutch CBS Statistics Netherlands http://statline.cbs.nl/Statweb/?LA=en

The data you select should meet the following requirements:

- n > 50
- at least two independent variables
- dichotomous a dependent variable
- Independence of observations

Conduct the following analysis on the data set:

1. Make a conceptual model underlying this research question
2. Conduct a logistic regression, examine whether adding individual indicators in the model improves the model compared to Null model. Make a final model with only significant predictor(s). For this model, calculate the pseudo R-square. Calculate the odd ratio for the predictors and their confidence interval
3. Make a crosstable of the predicted and observed response
4. Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2 and 3, and explain the conclusions that can be drawn.
5. Include annotated R script in the appendix of the report.

Part 3 – Multilevel model

For this part of the assignment you need to use the file set[x].cvs. To determine the version your group has to select, add up the student ID number from the group members and take modulo 3 of this number. The file includes longitudinal data collected from a large group of participants (Subjects) that in multiple sessions (session) completed a learning exercise for which exercise score (score) was collected. Note that the number of exercises completed between participants varies. Conduct a multilevel analysis to see whether over sessions the exercise score systematically vary. Besides a baseline model, create a model that includes session as a fixed factor, and uses a random intercept for the participants. Give an interpretation of the results and report the statistical results in a small paragraph for scientific publication.

Conduct the following analysis

1. Use graphics to inspect the distribution of the score, and relationship between session and score
2. Conduct multilevel analysis and calculate 95% confidence intervals, determine:
    a. If session has impact on people score
    b. If there is significant variance between the participants in their score
3. Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2 and 3, and explain the conclusions that can be drawn.
4. Include annotated R script in the appendix of the report