COURSEWORK ASSIGNMENT A - 2021

CS4125 – SEMINAR RESEARCH METHODOLOGY FOR DATA SCIENCE

For this coursework use markdown template file (Markdown report template assignment A). Submit the markdown file and knitted output pdf file. This file should include your answer, r code chunks, and output of the analyses. Be precise and brief in your answers.

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 markdown template file 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, and provide interpretation
3) Graphically examine the mean and the distribution of tweet's sentiments for each celebrity, and provide interpretation
4) Frequentist approach
    a) Compare linear models to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets (AICc, F-value, p-value etc). Provide brief interpretation
    b) 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. Provide brief interpretation of the results.
    c) Write a small section for a scientific publication, in which you report the results of the analyses of point 2-4, and explain the conclusions that can be drawn.
5) Bayesian approach
    a) Describe the mathematical model that you fit on the data. Take for this the most complete model that you fit on the data. Also explain your selection for the priors.
    b) Compare linear models to analyze the impact for adding information about the celebrity in explaining the sentiments of the tweets (e.g., WAIC, and 95% credibility interval of coefficients for individual celebrity). Provide brief interpretation of the results.
    c) Statistically compare sentiments of each celebrity pair. Provide brief interpretation of the results

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. Visually inspect if the variable page visits deviates from normal distribution, and discuss implication for general linear model analysis
4. Frequentist approach
    a. 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 (AICc, Chi-square, p-value etc). Provide brief interpretation of the results
    b. If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail. Provide brief interpretation of the results.
    c. Write a small section for a scientific publication, in which you report the results of the analyses of point 2-4, and explain the conclusions that can be drawn.
5. Bayesian approach
    a. Describe the mathematical model that you fit on the data. Take for this the most complete model that you fit on the data. Also explain your selection for the priors.
    b. Compare models to analyze to examine the added values of adding 2 factors and interaction between the factors in the model to predict page visits (e.g. WAIC, and 95% credibility interval of coefficients). Provide brief interpretation of the analysis results

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. Frequentist Approach

   a. Conduct multilevel analysis (AIC, Chi-square, p-values) and calculate 95% confidence intervals, determine:
      i. If session has impact on people score
      ii. If there is significant variance between the participants in their score
   b. 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.
3. Bayesian approach
   a. Describe the mathematical model that you fit on the data. Take for this the most complete model that you fit on the data (i.e., the model of point iii). Also explain your selection for the priors
   b. For the analysis only use subset of the first 100 participants (tips: add 1 to Subject id number). Compare the following models (WAIC), and provide brief interpretation of results:
      i. Model with only fixed intercept
      ii. Model extended with an adaptive prior for Subject id
      iii. Model extended session as a with fixed factor
   c. Examine the estimates of the model with the best fit (e.g., 95% credibility interval), and provide brief interpretation of the results.