COURSEWORK ASSIGNMENT A – 2022-2023

CS4125 – SEMINAR RESEARCH METHODOLOGY FOR DATA SCIENCE

For this coursework use markdown template file (Markdown report template assignment A 2023Q4). Submit the markdown file and knitted output pdf file. This file should include your answer, r code chunks, and relevant output of the analyses. Be precise and brief in your answers. Unnecessary long reports with long output are not appreciated. Also, you will use this report as the basis for group presentations. For the Bayesian analyses, use the rethinking and/or BayesianFirstAid library.

Part 1 – Design and set-up of true experiment

Write a plan for conducting an experiment on group of human test subjects. The plan should include the following items.

1) The motivation for the planned research. (Max 200 words)
2) The theory underlying the research. (Max 200 words) Preferable based on theories reported in the literature
3) Research questions that will be examined in an experiment (or alternatively, the hypothesis that will be tested in an experiment)
4) The related conceptual model, this model should include:
   a) Independent variable(s)
   b) Dependent variable
   c) Mediating variable (at least 1)
   d) Moderating variable (at least 1)
5) Experimental Design (the study should have a true experimental design to test a single hypothesis that, for simplicity, includes only independent variable(s) and dependent variable(s). In other words, mediating and moderating variables are not included in the experimental design)
6) Experimental procedure (how the experiment will be executed step by step)
7) Measures
8) Participants
9) Suggested statistical analyses

Part 2 – Generalized linear models

Question 1 Twitter sentiment analysis (Between groups – single factor)

Analyzing Twitter tweets about a specific celebrity, it is possible to get an overall sense of 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 three famous individuals. The markdown template file shows how you can obtain tweets automatically. This program uses the following file, which you need to place in your 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 figure of the conceptual model for the following research question: Is there a difference in the sentiment of the tweets related to the different individual?
2) 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. Assume a Gaussian distribution for the tweet's sentiment rating.
3) Create a synthetic data set with a clear difference between tweets' sentiments of celebrities for verifying your analysis later on. Report the values of the coefficients of the linear model used to generate synthetic data.
4) Graphically examine the mean and the distribution of tweet sentiments for each individual, and provide a short interpretation.
5) Frequentist approach. The aim is to conduct frequentist analysis by comparing linear model**s** to analyze whether the knowledge to which a celebrity tweet relates has a significant impact on explaining the sentiments of the tweets. Assume a Gaussian distribution for the tweet's sentiment rating.
   a) Verify your model analysis with synthetic data and show that it can reproduce the coefficients of the linear model that you used to generate the synthetic data set. Provide a short interpretation of the results, with a reflection of AICc, F-value, p-value etc.
   b) Redo the analysis now on the real tweet data set. Provide a short interpretation of the results, with an interpretation of AICc, F-value, p-value, etc.
   c) If a model that includes the individual better explains the sentiments of tweets than a model without such predictor, conduct a posthoc analysis with, e.g., Bonferroni correction to examine which celebrity tweets differ from the other individual's tweets. Provide a brief interpretation of the results.
   d) Write a small section for a scientific publication in which you report the results of the analyses of points 5b-5c, and explain the conclusions that can be drawn.
6) Bayesian approach. The aim is to conduct a Bayesian analysis to compare linear model**s** to analyze the impact of adding information about the individual in explaining the sentiments of the tweets (e.g., WAIC, and 95% credibility interval of coefficients for an individual celebrity).
   a) Verify your model analysis with synthetic data and show that it can reproduce the coefficients of the linear model that you used to

generate the synthetic data set. Provide a short interpretation of the results, with a reflection of WAIC, and 95% credibility interval of coefficients for individual celebrities.

b) Redo the analysis on the actual tweet data set. Provide a short interpretation of the results, with a reflection of WAIC, and 95% credibility interval of coefficients for individual celebrities.

c) Compare the sentiments per celebrity pair.

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

For this question, you must 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 must 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 must 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 a combination of the two had an impact on the number of pages visited.

1. Make a figure of the conceptual model underlying this research question.
2. Describe the mathematical model that you fit on the data. Take for this the complete model that you fit on the data. Also explain your selection for the priors. Assume Gaussian distribution for the number of page visits.
3. Create a synthetic data set with a clear interaction effect between the two factors for verifying your analysis later on. Report the values of the coefficients of the linear model used to generate synthetic data.
4. Graphically examine the **mean** page visits for the four different conditions. Give a short explanation of the figure.
5. Frequentist approach. The aim is to conduct frequentist analysis by comparing models to examine the added values of adding the two factors and the interaction between the factors in the model to predict page visits (AICc, Chi-square, p-value etc).
    a. Verify your model analysis with synthetic data and show that it can reproduce the coefficients of the linear model that you used to generate the synthetic data set. Provide a short interpretation of the results, with a reflection of AICc, F-value, p-value etc.
    b. Redo the analysis now on the real data set. Assume Gaussian distribution for the number of page visits. Provide a short interpretation of the results, with an interpretation of AICc, F-value, p-value, etc.
    c. Redo the analysis on the real tweet data set. This time assume a **Poisson** distribution for the number of page visits. For the best fitting models, examine graphically the distribution of the residuals for the model that assumes Gaussian distribution and

the model that assumes Poisson distribution. Give a brief interpretation of Poisson and Gaussian distribution assumptions.

   d. Continue with the model that assumes a Poisson distribution. If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail. Provide a brief interpretation of the results.
   e. Write a small section for a scientific publication in which you report the results of the analyses of points 5c-5e, and explain the conclusions that can be drawn.

6. Bayesian approach. The aim is to conduct a Bayesian analysis to compare models to analyze to examine the added values of adding two factors and the interaction between the factors in the model to predict page visits (e.g. WAIC, and 95% credibility interval of coefficients).
   a. Verify your model analysis with synthetic data and show that it can reproduce the coefficients of the linear model that you used to generate the synthetic data set. Provide a short interpretation of the results, with a reflection of WAIC, and 95% credibility interval of coefficients for individual celebrities.
   b. 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. This time, assume Poisson distribution for the number of page visits.
   c. Redo the analysis on actual data. Assume Poisson distribution for the number of page visits. Provide brief interpretation of the analysis results (e.g. WAIC, and 95% credibility interval of coefficients).

Part 3 – Multilevel model

For this part of the assignment, you must use the file set[x].csv. To determine the version your group must 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 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 varies. 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. Give a short description of the figure.
2. Frequentist Approach
   a. Conduct multilevel analysis (AIC, Chi-square, p-values) and calculate 95% confidence intervals, assume a Gaussian distribution for the scores, determine:
      i. If the session has an impact on people's score

ii. If there is a 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 points 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). Assume a Gaussian distribution for the scores. Also, explain your selection for the priors

b. Compare the following models (WAIC), and provide a brief interpretation of the results:

i. A model with only fixed intercept

ii. Model extended with an adaptive prior for Subject id

iii. Model extended session as a fixed factor

c. Examine the estimates of the model with the best fit (e.g., 95% credibility interval), and provide a brief interpretation of the results.