Table of Contents

[Learning outcomes being assessed through the assignment 1](#_Toc94365245)

[Description of the assignment 1](#_Toc94365246)

[What specific questions are being asked to the students as a part of this assignment? 2](#_Toc94365247)

[Submission requirements acceptable format of the assignment 3](#_Toc94365248)

[Deadline for submission 3](#_Toc94365249)

[Grading rubric being used to assess the assignment 4](#_Toc94365250)

# Learning outcomes being assessed through the assignment

The following module learning outcomes will be assessed through this assignment:

* To know when and how to fit simple/multiple Poisson regression models and interpret associated results
* To distill relevant and substantive numerical information and present it clearly and succinctly

# Description of the assignment

Below you find a description of the dataset (*drinking\_assignment.txt*) you need to analyse and associated questions to be answered (1 to 9).

**DATASET – Alcohol consumption in older adults**

The drinking dataset entails a sample of N=1419 participants between the ages of forty-five and seventy years old, who were sampled from general practitioners' databases. They all consumed alcohol in varying degrees. The file entails information on their alcohol consumption (AC), sex and smoking status. Your task consists in investigating the links between the available explanatory variables and alcohol consumption, with particular focus on smoking.

**Variables operationalisations:**

**Outcome:**

* **Alcohol consumption** was reported as a quantity and frequency measure for each drink type. Frequency was measured on a categorical scale, ranging from daily to never with six intermediate steps. These frequency data were converted to an average number of drinking occasions per week. Quantity is reported as the number of glasses drunk on a typical day at which the specific beverage was consumed. Based on the reported frequency and quantity measures the variable Total weekly AC volume (glasses/week) was computed (quantity x frequency).

**Explanatory variables:**

- **Sex** was coded ‘0’ for females and ‘1’ for males

- **Smoking status** was operationalized as pack years, which is a measure of lifetime tobacco exposure based on the number of years a participant had smoked and the average number of smokes per day.

# What specific questions are being asked to the students as a part of this assignment?

**Questions:**

1. Run an Exploratory data analysis (EDA) to get preliminary impressions about the associations (report associated plots to substantiate your answers). Which associations do you expect, and in which directions?
2. Smoking is a count variable. Can you add it to the model in its original scale, i.e. does linearity/monotonicity hold for both variables? Motivate your answer;
3. Explain why the use of the **linear regression** models may not be the adequate choice to analyse this dataset;
4. Run **simple** Poisson regression models with all explanatory variables, one at a time. Are they significantly linked to the outcome? Interpret the associated effects;
5. Run a **multiple** model with all Xs (mutually adjusted for each other). Write down the estimated multiple model in its full extension (the equation);
6. Draw conclusions regarding observed associations, without forgetting to interpret them one more time. Are there differences between the **crude** (question 4) and **adjusted** (question 5) associations? If yes, what would be the possible explanations, i.e. what is the third variable problem in the data, **confounding** or **interaction**? (Please note that in the latter case, the adjusted model needs to be extended to consider an interaction term);
7. Once selecting the final model, check its goodness of fit (GOF). Did you encounter any over-dispersion problem? Motivate your answer;
8. If yes (question 7), re-run the model again, this time using the ‘*quasipoisson*’ family. Are the conclusions the same as the ones from “*Poisson*” model?
9. Write down a brief paragraph describing the objective of your investigation and main conclusions based on your analyses (no statistical jargon, please- just a summary of the results in plain and accessible language);

# Submission requirements acceptable format of the assignment

**Do not** provide R codes together with the answers of your questions. Add them (screen shots of the fitted models and associated outputs) at the end of the document (a MS word file) as an appendix. Please do so in structured and ordered fashion, i.e. with associated questions’ numbers. If the questions require some visual displays, add the graphs to the answer (not to the appendix).

# 

# Deadline for submission

The deadline for submitting this assignment is 15th of March 2022.

# Grading rubric being used to assess the assignment

Your assignment will be grading according to the following scoring system

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Grade** | **between 0-3** | **between 4-5** | **between 6-8** | **between 9-10** | |
|  | **Absent - Poor** | **Weak** | **Good-V Good** | **Excellent** | |
| Question 1 – choose and present proper graphical displays with associated interpretation  Question 2 – provide a correct answer with proper motivation (and evidence)  Question 3 - address relevant features in the linear vs. Poisson regression comparison, with explanations  Question 4 – provide relevant numerical information and associated interpretation  Question 5 – write down the correct algebraic expression of the final model, with obtained regression coefficients’ estimates  Question 6 – provide a well-reasoned and correct answer/speculation  Question 7 – provide and interpret relevant statistics, with motivation  Question 8 – provide a correct answer with motivation  Question 9 – present a brief summary of the results in a condensed, informative accessible paragraph (**20 points**) |  |  |  |  | |
| **Grades**   |  |  | | --- | --- | | **0-40** | The student made little or no attempt to complete the assignment. | | **40-50** | Some effort had been made, but answers/solutions were inadequately reported. No explanation was provided in a shot paragraph with each solution. | | **50-60** | Two questions were correctly answered, but comments were missing and some of the results are not correct. No explanation was provided in a shot paragraph with each solution. | | **60-70** | Three questions were correctly answered but comments were missing and some of the results are not correct. Missing or incorrect explanation was provided in a shot paragraph with each solution | | **70-80** | All four questions were correctly coded but comments were missing and some of the results are not correct. Missing or incomplete explanation was provided in a shot paragraph with each solution | | **80-100** | All nine questions were correctly answered with well-reasoned comments. All the reported results were correct and proper explanation was provided in a shot paragraph with each solution | | | | | |  | |  |  |  |
|  | | | | |  | | | | |
|  | | | | |  | | | | |