PUBH 7430- Assignment 1

On this assignment, as in all assignments, **inclusion of screenshots of R/SAS output or R markdown raw output is not acceptable**. *Tables and plots prepared from statistical output should be appropriate for inclusion in a scientific report, with all statistics rounded to a reasonable number of significant digits.* **Obtaining the correct results from an analysis is only a portion of each exercise**. *It is also very important to write clearly about what your results mean. This includes presenting your results in a way that clearly answers the question and places little burden on the reader.*

You should submit your assignment on Canvas as a PDF file, titled "PubH7430_LastNameFirstName_A1.pdf" with your own name replacing LastNameFirstName. While no code should be included in the body of the assignment, you **must include your code as an appendix** *at the end of the assignment. Keep in mind that while working together on homework assignments is permitted, each student is expected to independently write up homework assignments, including any code, in their own words.*

Background

Each year the U.S. Naval Postgraduate School sets aside a "Discovery Day" during which the general public is invited into their laboratories. This dataset is adapted from data collected on October 21st 1995, when visitors could test their reaction times and hand-eye coordination in the Human Systems Integration Laboratory. The variable of interest, "anticipatory timing", was measured by a Bassin timer, which measures a person's ability to estimate the speed of a moving light and its arrival at a designated point. The timer consists of a 10 foot row of lights which is controlled by a variable speed potentiometer. The lights are switched on sequentially from one end to the other so that light "travels" at 5 miles per hour down the timer. Each visitor was instructed to anticipate the "arrival" of the light at one end of the timer and at that time to swing a plastic bat across a light beam at the same end of the timer. An automatic timing device measured the difference between the breaking of the beam and the actual arrival of the light. In the original data, a negative time value for a trial indicated that the bat broke the beam before the light actually arrived; in the version provided to you, all times have been transformed to positive values, so that the values reflect the magnitude of how far off the participant was in timing.

Each of 113 visitors completed the trial five times. Age and gender were also recorded, since the researchers were interested in age and gender differences in reaction times. Visitors tended to come in family groups, but that information was not recorded. The final dataset contains data from 107 individuals after excluding observations from those under age 6 (who may not have fully understood the task).

You can find these data from the file *timetrial.csv* on Canvas. These data are organized in wide format, with one row per person and one column for each of the five trials. Depending on what software you use for plotting, they may need to be reshaped to long format, i.e. to have one row per trial instead.

Questions

1. **Notation. [7 pts]** Consider the notation presented in Lecture 5, where outcomes are denoted by $\boldsymbol{Y}$ and covariates/predictors by $\boldsymbol{X}$. Suppose we are interested in modeling the response times in this dataset as a function of age (treated as a continuous variable), gender (coded as Male = 0, Female = 1), and trial number (treated as a continuous variable), via the simple linear model

$$\boldsymbol{Y} = \boldsymbol{\beta}_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

For this model as applied to the time-trial data:

(a) [1 pt] Write down the vector $\boldsymbol{Y}_{20}$ corresponding to the outcomes of the twentieth subject in the dataset

(b) [1 pt] What is the length of the full vector of responses $\boldsymbol{Y}$?

(c) [1 pt] Compute the correlation matrix, *Corr*($\boldsymbol{Y}_{20}$), assuming that for every trial, j, *Var(*Y_{20j}*)*=0.05, and for $j \neq k$ *Cov(*Y_{20j}, Y_{20k}*)*=0.02.

(d) [2 pt] What is the dimension of the matrix $\Sigma = Var(\boldsymbol{Y})$?

(e) [1 pt] Write down the matrix $\boldsymbol{X}_{20}$ corresponding to the matrix of covariates for the twentieth subject (Id=20) in the dataset.

(f) [1 pt] What is the dimension of the full matrix of covariates $\boldsymbol{X}$?

2. **Exploratory analysis - changes across trials. [11 pts]** Using the plots and summary statistics discussed in class, describe any trends in the timings across the five trials. For each plot/table, comment in one or two sentences on the type of information the plot/table tells you. Note that your goal is to perform exploratory data analysis (EDA) in this question, so there is no need to perform statistical tests. In particular, your EDA should address the following:

(a) [3 pts] Are there any outliers (either individual observations or trajectories) that you might consider excluding from your analyses? If yes, explain your answer.

(b) [4 pts] Are visitors in general improving across the five trials? What pattern, if any does this improvement follow?

(c) [4 pts] Do trends in the outcome across trials differ between kids (age<18) and adults (age $\geq$ 18)?

3. **Exploratory analysis - correlation structure [6 pts].**

(a) [4 pts] Using plots and/or summary statistics describe how the correlation between response times changes across trials.

(b) [2 pts] What "named" correlation structure (e.g. independence, exchangeable, AR-1, Toeplitz, unstructured) appears to be most consistent with the data? Justify your response.

4. **Estimating effects.[16 pts]**

(a) [4 pts] Suppose we are interested in testing whether there is a "learning" effect across trials. Perform a t-test to determine if subjects were more accurate during the fifth trial than the first trial. Report the results in the form of a sentence or two suitable for publication in a scientific journal. Be sure to specify the kind of t-test performed and why you choose that test.

(b) Now suppose that we wish to test whether adults have different response times than kids.

  i. [4 pts] Perform five separate t-tests to determine if kids had a different mean response time than adults on each of the five trials. Report your findings in a table or figure and summarize the results in a few sentences. Be sure to specify the kind of t-test performed and explain why this is the appropriate choice.

  ii. [4 pts] Perform a single t-test to compare the response times of kids and adults. To do this you will need to calculate a measure for each individual which summarizes their data from all five trials. Summarize your results in a sentence or two.

  iii. [4 pts] Briefly summarize the advantages and disadvantages of the analyses proposed in parts 4(b)i and 4(b)ii.