Experimental Design for Data Science

INF2178 Technical Assignment 3

Due: March 23rd at 11:59pm on GitHub (**in the Assignment 3 folder**)
Note: You **MUST** name your files in the format FIRSTNAME-LASTNAME-A3.ipynb and FIRSTNAME-LASTNAME-A3.pdf (example: VICTORIA-CHUI-A3.pdf). You will be penalized if you do not.

In this assignment, you will be exploring a dataset and performing analysis using **one-way ANCOVAs**. We are looking for knowledge of quantitative analysis through the use of ANCOVAs in Python, and a theoretical understanding of their use through the write-up. You will need to **submit both a .ipynb** Jupyter notebook (with comments throughout) **and a .pdf** (with a short write-up/narrative). To submit these on GitHub, please make a pull request at the appropriate location, following the instructions in the GitHub reference video on Quercus.

Data

A subset of data from an early child longitudinal study (1998-99) will be used in this assignment. The data has reading, math, and general knowledge scores for fall 1998 and spring 1999 measurements, evaluating Kindergarten students over the span of several months. Income category is also given (the only categorical variable).

In this assignment, you will examine Kindergarten scores by using the dataset titled INF2178_A3_data.csv. The first six columns of the data are continuous variables, representing the fall/spring reading, math, and general knowledge scores of individual students. Income is given as a continuous variable, although is only used to derive the income group variable. You are welcome to use general knowledge as a baseline, to compare how students' reading and math scores change over time, by income group.

Instructions

1. Examine the dataset. Not every column is going to be useful for you in this assignment; however, you are welcome to utilize every column if you are interested.
2. **Python**: Perform quantitative analysis using **multiple (as many as you feel necessary) one-way ANCOVAs**. Utilizing your knowledge of ANCOVAs and the code from class, perform quantitative analysis as you see fit.
    a. Create a narrative, tell us something about the data. What story could you tell from this preliminary analysis, or what further analysis would you need to do to explore the research question(s) you have in mind?
    b. In your code, you need to perform exploratory data analysis, run the model(s), create interaction plots, and test the assumptions for running a one-way ANCOVA to earn the maximum points.
3. **PDF**: Write a short narrative (no longer than 6 pages IN TOTAL - including figures and tables) explaining your process and what you learned from the data.
    a. Include headings and sub sectioning as needed.

b. Your figures/tables should be professional. Tip: **You do not want to provide screenshots of your code/output/data frames** in your write up and you do not need to mention the various functions or Python libraries you used to conduct your analyses. We will see this from your code.
c. Your write up needs to include research questions, exploratory data analysis, interaction plots, results of testing the assumptions for running a one-way ANCOVA, and ANCOVA results for maximum points.

Marks Breakdown

This assignment is worth 20% of your final grade. The grading will be broken down as follows:

% of Assignment	**Item**
40%	Functionality of code and use of appropriate code (we run your code)
10%	Code comments
40%	Narrative of findings (**research questions**, results, discussion…)
10%	Successful submission to GitHub and naming of files

IMPORTANT Feedback from A1 to Improve A3

Check your assignment 3 write-up for these common errors **before submission**

1. P-value cannot be 0, instead should be stated as < 0.001.
2. What were your research questions and did your analysis answer them?
3. Figure text (legends/axis labels) should be in large enough font and nonoverlapping. Use figure and table numbering.
4. Dataframes, summary statistics, and t-stat/p-value should be in properly formatted tables instead of screenshots. Round your p-value/t-stat to 2 or 3 significant figures.
5. Instead of using verbatim column names, use a meaningful phrase to represent what the variable means.
6. Put your name on your write-up pdf and code notebook!!!!
7. Use the dataset that was uploaded to Quercus when running any code analysis. I.e. if we provide a csv, run a csv file. **Do not create your own excel file**. We grade your code by loading the dataset we provide and running all your code, so if we cannot run your code then you will lose marks. A good step to take before submission is to restart your kernel and run all cells to check you have not missed defining variable names or packages etc.