BIOS 525 - 2020 Midterm Data Analysis Project

Instructions:

1. Submit report by Monday 1pm October 5th.
2. 4-page maximum, single-spaced, Arial font size 11, 1-in margins, excluding tables and figures.
3. Include *all* analytic code in an appendix.
4. Do not include any output from statistical software directly in the report.
5. All work must be COMPLETELY independent.
6. Email all questions to the instructor. Aggregated responses will be sent to all students.

Background The dataset (*Midterm2020_Data.RData)* contains survey data from 3,600 children in the Early Childhood Longitudinal Survey carried out by the Institute of Education Sciences (http://nces.ed.gov/ecls/kindergarten.asp). This national survey began in 1998 where children in kindergarten were recruited and then followed through the 8th grade. There were 7 repeated measurements taken at: fall and spring of kindergarten, fall and spring of 1st grade, spring of 3rd grade, spring of 5th grad and spring of 8th grade. This extensive dataset has been an important data source for education and social research. Here you will analyze only a subset of the participants.

Variable Codebook:

1. *race* — Participant race
   - 1 = White, non-Hispanic
   - 2 = Black or African American, non-Hispanic
   - 3 = Hispanic
   - 5 = Asian
2. *gender* — Participant gender
   - 1 = male
   - 2 = female
3. *dad_edu* — Highest education attainment by the father
   - less HS = did not complete high school
   - HS or GED = high school diploma or GED
   - College or AD = Associate's or Bachelor's degree
   - Graduate or Prof = Graduate or professional degree
4. *age* — Participant's age in months at baseline
5. *t1, t2, …, t7* — Math score (unit-less with range: 0 to 300) for period 1, to 7.

PART I. Exploratory Data Analysis (20 points; suggested time = 2 hours)

1. Understand the clustering structure of the data.
2. Examine relevant summary statistics for the sample.
3. Examine the missing data pattern in the outcome variables *t1*, *t2*, …, *t7*.
4. Examine the **univariate association** between *math scores* and the child's age. Is there evidence of non-linear relationship? Does the association vary by gender, race, and father's education attainment?

5. Examine model fit. Are there outliners? Should the response variable be transformed?

Summarize your findings in paragraphs (3 max) with optional tables/figures (2 max).

PART II. Modeling Math Score Trajectories (50 points; suggested time =3 hours)

Based on your findings from Part I, fit **random intercept models or marginal models** to examine the influence of race, gender, and father's education on a child's math ability longitudinally.

- Clearly state your scientific questions. Describe you approach and write down the corresponding statistical model (4 paragraphs max).
- Summarize your findings (4 paragraphs max) with optional tables/figures.

PART III. Discussion (30 points)

Discuss some limitations of your analysis. Examples include model specification, missing data, missing variables, confounding, and HOW these may impact your findings. (2 paragraphs max)

Discuss why your choice of random intercept or marginal model is appropriate here, as well as *other* modeling choices you made in Part II (2 paragraphs max).

Tips for Success!

1. Start early.
2. Don't underestimate the amount of time needed for writing (suggested time = 3 hours).
3. These are *real data* and they are going to be messy.
4. There is no *true model*, but some models can be useful. Everyone's model and results will be slightly different. Make sure you justify your modeling decisions.
5. When unsure about whether you should do something, consider the following.
    - Does it add to your story? Can you address it or is it a limitation worth discussing?
    - Try it out with simple analyses.
    - Make a decision. Write a sentence or two. Then move on.
6. Make sure all reported quantitative estimates have an interval estimate or a standard error when appropriate.
7. Make sure your subscripts are correct when writing down the model.
8. Make sure units are added when appropriate.
9. Make sure axis labels and legends are included in figures.
10. Avoid long and complex sentences.