Stat 313: Linear Regression Project

Write-Up Guidelines

You will be summarizing your results in a (guided) written report following the **Project Writing Guidelines** posted on Canvas. The results **must** be written in the RMarkdown template provided. You can include all pertinent plots inline in the typed document, rather than in the appendix. Your group will be submitting both the RMarkdown file used to generate your report and the knitted HTML file to Canvas by **Sunday, february 7 at 11:59pm**.

Begin by downloading the RMarkdown and data files from Canvas and saving them on your computer in the **same** folder. Then open the RMarkdown file in RStudio and run the first code chunks loading packages and reading the data in. If you encounter an error, make sure that the data are located in the **same** folder as the RMarkdown file!

Data Background

Today, most colleges and universities require a standardized test for applicants, such as the SAT or the ACT. Due to this requirement, roughly 2 million students in the U.S. take the SAT each year. Disappointingly, the Varsity Blues scandal has left a large number of people questioning the role that wealth plays in student's standardized test scores.

A 2013 study by Dixon-Roman and Mcardle titled "Race, Poverty, and SAT Scores" found that "the effects of family income on SAT scores, though relatively modest in contrasts to high school achievement, are substantial. These researchers investigated the relationship between high school achievement and SAT scores, controlling for mother's and father's education and family income. Specifically, this study considered SAT data from 2003.

In this project we will investigate the relationship between the average SAT math score and average SAT verbal score. We will then explore if this relationship differs for different family incomes. We will use data for each state in the United States from 2005 to 2015, but **will not** use state or year as explanatory variables in our model.

As is typical in educational data, students are not randomly assigned to a "treatment" group, rather the characteristics of thes students are observed. Additionally, as these are data from **every** student who took the SAT exam in each state from 2005 to 2015, we have a *census* of the population.

Introduction (5 pts)

- (2 pts) Give a brief background of the research problem and how the data were collected. Make sure to describe who the unit of study is (e.g. students? states? years?)!
- (3 pts) Clearly outline the question(s) of interest you will address with the statistical analysis. The more specific you define the question of interest here, the easier the rest of the analysis and report will be. The research questions should start with, "What is the relationship between. . . " and should be as specific as possible. Your *Summary of Statistical Findings* should directly answer the question(s) you pose here.

Statistical Methods (15 pts)

This section should lay out the steps, decisions, and logic leading to the statistical model you will use to answer the research question of interest.

- (2 pts) Describe the response and explanatory variables.
- (2 pts) Provide a summary table of the median average SAT math and SAT verbal scores for every year in the data

Note: Keep in mind that the `Math` and `Verbal` variables represent the *average* SAT math and SAT verbal scores.

- (3 pts) Produce data visualizations exploring the relationship(s) you are interested in investigating. Describe what you see in the visualizations, making direct references to the plots!

Note: Keep in mind that there are a large number of observations in the dataset, so you may want to use tools that help to alleviate overplotting!

- (4 pts) Describe the appropriate statistical model you will use to answer the question(s) of interest that you stated previously. Be specific about why the method being used are appropriate for the investigation at hand (e.g. types of variables).
- (4 pts) Check all model conditions of the statistical method you used. Describe what each condition is **in the context of these data**. Reference and include appropriate plots necessary for checking the model conditions in line. What are your conclusions regarding the conditions? Justify your conclusions!

Summary of Statistical Findings (10 pts)

In this section you will write up your findings for each research question of interest.

- (5 pts) What is your conclusion for the questions of interest? Namely, "What is the relationship between the SAT verbal and SAT math scores?" and "How does the relationship differ by family income?". Base your conclusion on the visualizations you created **and** the regression model you found.
- (5 pts) Interpret **in the context of the data** each of the coefficient estimates you got from R.

Scope of Inference (4 pts)

Write a brief Scope of Inference statement. Specifically, answer these two questions and comment on their implications:

- (2 pts) Were the observations randomly selected from some larger population? Based on the sampling method used, what larger population can you infer the results to?
- (2 pts) Was the explanatory variable randomly assigned to observations? Based on the study design, are cause-and-effect statements justified?

Make sure you write the scope of inference specific to the language of the data (not just generic statements)!

Project Presentation (3 pts)

- (1 pts) Your report should not have any spelling errors! To check for spelling errors in RStudio, click the green check mark button next to the "Knit" button.
- (2 pts) Your report should look as neat and professional as possible. Make sure that your figures don't end up in the middle of your paragraphs, and that your sections have headings. If you would like to fine tune the appearance of your report, please post questions to Canvas and I will respond ASAP.

Note on Figures: I expect that the figures are included in the section they are discussed not at the end of the report.

Note on Model Output: Please try to make the output from the statistical model look as nice as possible. Output from `summary()` looks terrible in a statistical report! Try using the `get_regression_table()` function from the **moderndive** package or the `tidy()` function from the **broom** package.

Group Evaluation (3 pts)

Each member of your group will fill out a group evaluation form detailing each member's contributions, cooperation, communication, and participation.