STAT 7630 - Data Analysis Project

D. Pluta

Spring 2024

General Info

- The purpose of this project is for you to demonstrate the application of linear regression on a realistic data set following the procedures and methods learned in class.
- Due: Monday, May 6th by 23:59 on D2L

Report Requirements

- The submitted report must be no more than 20 pages, including all figures, tables, and appendices. (This page limit does not include R code.)
- The report should be organized with the following sections:
    - Introduction
        - Describe the overall context of the project or problem setting
        - Introduce the specific problem you're interested in, and succinctly explain why it's an interesting research question
        - Give a brief summary and review of previous related work
        - Motivate your study as an improvement on or addition to the existing literature
        - Give a 1 paragraph description of the structure of the paper.
    - Materials and Methods
        - Description of data
        - Description of experimental design and data collection procedures
        - Precise statistical formulation of the scientific questions of interest and the statistical methods that were applied
    - Results
        - Presentation of main data analysis results, organized according to the scientific questions of interest posed in the methods section
        - Brief presentation of main speculative modeling results (if interesting)
    - Discussion
        - Summarize the main findings of the paper.
        - Comment on any limitations or drawbacks of the current study.
        - Provide a brief discussion of speculative modeling results and possible directions for future studies.
    - Appendix
        - Details of methods applied
        - Diagnostics of main model, and description of any remediations that were required
        - Further details and results of other speculative models tried
        - Additional Tables
        - Additional Figures

Steps

1. Pick an area or topic of interest. Some examples:

- Healthcare/Medical
- Education and Educational Outcomes
- Fundamental science (Physics, Chemistry, Biology)
- Economic (stock market, inflation, import/export relationships)
- Sociological/Political

2. Decide on one or two question(s) of interest and find a relevant data set.
3. Formulate the question(s) statistically, and propose a linear regression model to answer the question(s).
4. Exploratory data analysis
5. Modeling
6. Diagnostics
7. Interpretation
8. Exploratory/Speculative modeling
9. Summary of results and interpretation in terms of the applied question(s) of interest.

Minimum Expected Content

- Introduction that explains the context of the problem, and the motivation for the question you are posing
- Description of the dataset, e.g., where did it come from, how was it collected, why was it collected originally, etc.
- “Table 1” summary descriptions of data set distributions of response and covariates
- Explanation of adjustment variables included in the model (e.g., confounder, precision, or variable of interest).
- Table of regression results from main model, including confidence intervals of any coefficients of interest
- Clear and precise interpretation of analysis results in context of the question of interest
- One or two plots illustrating the analysis results
- Appendix with sufficient diagnostic analysis of main model
- Discussion of limitations and proposals for future studies