End-of-Semester Reflection

Daniel Posmik

2024-12-12

Project 1: Exploratory Data Analysis

For Project 1, my revisions focused on improving the presentation of results. I decided against conducting a complete reanalysis. Following the instructor's feedback, I made several changes to enhance clarity and accessibility for an audience that may not be familiar with the topic. One of the key adjustments was revising the abstract to make it more concise, aiminng at better introducing the study's objectives and findings. Further, I reorganized portions of the presentation to improve its flow. For instance, the original report included a large chart detailing the missingness patterns in the data, which I decided feels somewhat redundant. I decided to remove this chart, focusing instead on presenting the most essential information.

Beyond these changes, I also implemented several detailed suggestions from the instructor to refine the report further. One notable improvement was replacing R variable names with the corresponding real-world variable names. This adjustment makes the report more "flowy", allowing readers to cross-reference results more easily while also improving its overall readability and aesthetics. The most substantial improvements, however, involved enhancing the interpretations provided in the text. I worked to add more context and depth to the discussion of the results. For instance, when describing the scatter plot of age and completion time, I expanded the interpretation to include a more nuanced descriptive analysis. I would say this part also exemplifies my most important piece of growth. Rather than looking for trends that I am biased towards "wanting to see", a solid EDA carefully analyzes all trends in detail and with an objective lens.

I believe that this project was a fantastic opportunity to refine this skill, especially as a Ph.D. student. A lot of times, the focus is immediately on the modeling, and pre-existing biases may worsen an analysis and lead to incorrect conclusions. This project

was a great reminder of the importance of a thorough EDA and the importance of being objective in the analysis. Another critical aspect of my revisions addressed a data quality issue in the relative humidity variable. The data exhibited significant bunching at 0% relative humidity, likely caused by the sensor's inability to measure humidity below a specific threshold. I discussed this issue in the report, noting that additional contextual information about data provenance could help resolve such anomalies. For instance, if more details about the data collection methods were available, we could adjust or verify the quality of these variables. In the absence of this information, I proposed alternative solutions, such as removing the 0% values and imputing them with more reasonable estimates to maintain the integrity of the analysis.

These revisions collectively improve the report's coherence, usability, and interpretability, making it a stronger and more polished final product.

Project 2: Regression Analysis

A major focus of my revisions for Project 2 was around the justifications for methods used and improving overall report readability. I began with better justifying the choice of LASSO for variable selection and going into more detail on variable selection. Notably, a thought that wasn't well communicated initially were my concerns about overfitting. Initially, analysis shows that only the FTCD variable is a significant predictor of smoking cessation. I proceeded by grouping variables in the same category and visualizing their correlations. This includes the groups "demographic variables", "socioeconomic variables", "smoking variables", "addiction variables", and "mental health variables". Due to low n and high p, I felt like we needed to reduce the variables to adjust for drastically. Although we do not observe strong correlations, I feel like these "clusters" were chosen with enough reasoning to qualitatively justify the drastic reduction in control variables. Certainly, this is not an ideal setting, but neither is the sample size. Since we are already seeing a poor model fit, I hoped to avoid spurious results by reducing the number of predictors so drastically.

The choice of LASSO was to gather more evidence for or against the above reasoning. I should have done a better job explaining that I believe the logistic regression gets the job done well, but I did want to see whether adding the penalty term leads to shrinkage towards 0 for any of the variables chosen. This was not the case, although the cigarette reward value and age variables were shrunk significantly. On the contrary, if we had observed significant shrinkage across categories, this could have warranted even more reduction in the set of controls. The LASSO regression was used as a check and I have revised the report to better communicate this.

Another significant change was a more holistic discussion of the limitations. Incidentally, I feel like this is my most important take-away from this project. Throughout this project, I constantly felt like the limitations were compounding. Starting from the drastic reduction in the number of predictors, due to the low n, and ranging to the choice of interactions. This project can certainly be more exhaustive but given the constraints in time, choices must be made. I feel like communicating this better is key. Notably, I also feel like my choice of interactions was indeed well justified, however, I could have given more background on the limitations. I have now added a paragraph on this in the limitations section.

Overall Thoughts

I appreciate how this course has sharpened my ability to write and communicate statistical work. I appreciate the guidance and high expectations in terms of the presentation of results. One of my perceived weaknesses as a researcher is my ability to revise results after the initial "excitement" of the analysis has passed. Ironically, I currently feel this writing this report and preparing the relfection document/ finalizing the revisions. I hope that I can continue to build on this skill and that it will become more natural over time. I am grateful for the feedback and the opportunity to improve my work.