Final Reflection Essay on Projects Revisions

Wenyu Yang

Project1 Reflections

In my revised project1–"Impact of Age and Environmental Conditions on Marathon Performance," I took a comprehensive enhancement of the data analysis and presentation methodologies to improve their clarity and depth. Initially, the project relied primarily on simple visual plots which, while effective for initial observations, did not capture the complexity necessary for a nuanced understanding of the intricate datasets involved. Recognizing the limitations of these initial visualizations, it became apparent that a more thorough overhaul of data presentation techniques was necessary to better communicate the intricate relationships and patterns within the data.

Initially, I enhance the interpretability and depth of the data from my study by transitioning from raw data visuals to a more structured format. To achieve this, I used the tbl_summary function from the gtsummary package in R, which was important in organizing the data according to essential demographic and environmental variables, such as sex and race conditions. These variables are particularly critical in analyzing marathon performances, as they offer insights into how different groups and conditions affect athletic outcomes. The transition to a summary table format was a strategic move designed to present the data in a way that enhanced both readability and comprehensiveness. The table included detailed statistical summaries for continuous variables, such as mean, standard deviation, minimum, and maximum values, and for categorical variables, which provided counts and percentages. This structured presentation was a significant improvement over the initial visuals, facilitating a clearer and more accessible view of the data. It allowed for immediate comparison and contrast of the impacts of various factors, making it much easier for the audience—whether seasoned statisticians or casual readers—to grasp the complex interactions within the data.

In addition to improving data readability, I also enhance the visual representation of the trends in the marathon performance data. To this end, I added a trend line to the existing plots that segmented marathon performance by age and sex. This addition was implemented using a loess smoothing method, known for its flexibility and effectiveness in visualizing data trends without the constraints imposed by traditional parametric assumptions. The loess method, which utilizes locally weighted scatterplot smoothing, is particularly well-suited for

our analysis due to its capacity to model non-linear relationships that are frequently present in age-related performance data.

My method choice was in order to a more accurate patterns within the marathon data. Unlike linear models that may misrepresent real-world dynamics, the loess smoothing method allows for a more nuanced view of how performance metrics change across different age and genders. This approach not only highlights the trends but also illustrates the subtleties within the data, offering a richer perspective. The revisions were driven by specific analytical needs. The transformation to a summary table format was to present the data more coherently. This format allows for easier comparison and understanding of the impact of various factors like sex, age, and environmental conditions on marathon performance. By presenting data in a table format, readers can quickly ascertain the distribution of data, which is crucial for statistical analysis. Additionally, adding the trend line in the plot was to provide an intuitive understanding of how age and sex interact to influence marathon performance. This visuals help in highlighting the subtle variations in performance across different age groups and between sexes, thus emphasizing the differential impact of physiological factors on marathon outcomes.

When I was revising the project, I learned a few lessons. Firstly, the importance of data presentation was underscored, showing that transforming raw data into structured tables is important. Furthermore, enhancing visual plots with trend lines improved the interpretative value of the data. This highlighted how structured data could lead to more robust conclusions. Moreover, the trend line provided a clear visual representation of the data trends, which is important in understanding the relationships in a dataset. Visual aids like trend lines help in identifying patterns that are not immediately apparent through raw data analysis. In a nutshell, I learned importance of integrating visual elements in statistical reporting to enhance the clarity and impact of the findings.

Project2 Reflections

In my revised project2–"Analysis of Smoking Cessation Strategies and Psychological Influences in Major Depressive Disorder," I made extensive revision and focused on refining the statistical analysis to better align with the project's objectives. The initial approach incorporated a broad range of model comparison metrics, including AIC and BIC, which, while comprehensive, sometimes cluttered the analysis without providing substantial incremental insights. Therefore, the decision was made to streamline the process and enhance the analytical framework.

In my revision of the statistical analysis, I made adjustments to streamline and enhance the efficacy of our methodologies. Initially, the project used a variety of comparison metrics, including Akaike Information Criterion and Bayesian Information Criterion. While these metrics are commonly used for model selection, they often do not provide the most direct insight into model performance, particularly in applied health research settings where the interpretability and practical applicability of the models are crucial. To refine the analysis, I removed these

redundant metrics, shifting the focus towards more robust and directly informative techniques. K-fold cross-validation was prioritized as a key method for assessing model generalizability. This technique involves partitioning the data into a set number of subsets, and using one fold as the test set while the others form the training set. This process is repeated such that each fold serves as the test set once. K-fold cross-validation is instrumental in understanding how the model performs across different subsets of data, providing a comprehensive view of its reliability and stability across varied samples.

Additionally, I incorporated regularization methods like Lasso and Ridge regression. These techniques are valuable in handling multicollinearity and overfitting, common issues in complex datasets like those used in psychological and behavioral research. Lasso regression aids in feature selection by penalizing the absolute size of the regression coefficients, effectively reducing the number of variables included in the model by assigning zero to less important predictors. Ridge regression, on the other hand, penalizes the square of the coefficients, thus shrinking them towards zero but typically keeping all variables in the model. This approach allows us to maintain a broader dataset while controlling for multicollinearity. The practical application of Logistic Regression was emphasized due to its robust performance across several key metrics—accuracy, recall, precision, and F1 score. Logistic Regression proved especially adept at identifying true positives with a high degree of reliability while minimizing false positives, a critical factor in the context of our study where the correct identification of successful cessation cases could significantly impact subsequent therapeutic interventions and recommendations.

Moreover, I eliminated Forward Stepwise and Backward Stepwise from the analysis. Initially included as part of the exploratory data analysis phase, these methods were found to contribute little to our understanding of the dataset, showing zero specificity in our specific application. It was a strategic decision aimed at enhancing the efficiency and focus of our analysis, thereby streamlining the research process. I also included the variable "Var BA" across all models to ensure greater consistency in the analysis and to explore its potential influence on the outcomes more thoroughly. "Var BA" was hypothesized to be a significant predictor in the context of smoking cessation and was thus tested across different modeling approaches to ascertain its impact thoroughly.

My decision to remove less relevant metrics such as AIC and BIC was a strategic aimed at simplifying and focusing our study. These metrics, while useful in some contexts, did not contribute to our understanding of model performance or the applicability in the context of smoking cessation. Instead, the shift towards employing k-fold cross-validation and regularization methods such as Lasso and Ridge was designed to enhance the generalizability and robustness of our models. This offer insights into model performance allowed for a more streamlined and impactful analysis. For instance, k-fold cross-validation provided a thorough assessment of how our models performed across different subsets of data, thus ensuring that our findings were not merely artifacts of a particular sample but held broader applicability.

I also chose models that excelled in identifying true positives and minimizing false positives. I improved the accuracy and reliability of the results. Logistic Regression, with its clear

interpretive framework, allowed us to present things applicable to clinical settings. This made the research not only academically rigorous but also relevant, providing insights that could be directly utilized in the development of smoking cessation programs. It served as a profound learning experience, underscoring the critical importance of selecting appropriate modeling techniques and metrics that align closely with specific research objectives and constraints. The balance between model complexity and interpretability is a delicate one, particularly in fields where the implications of research can have direct clinical consequences. In our study, ensuring that our models were both statistically sound and clinically relevant was paramount. This required a careful consideration of each model's strengths and limitations, particularly in terms of their capacity to handle complexities such as class imbalance and missing data.

I learned that class imbalance and missing data are prevalent in studies and can significantly skew results if not properly managed. Our proactive approach in addressing these challenges through strategic model selection and data handling techniques highlighted the necessity of thorough methodological planning. By implementing techniques like oversampling or using advanced imputation methods for missing data, we improved the reliability and validity of our models, ensuring that they accurately reflected the true dynamics of smoking cessation among individuals with Major Depressive Disorder. In conclusion, revising this project was an educational experience for me to get insights into data analysis, statistical modeling, and the effective presentation of complex information. Reflecting on the feedback and making changes not only improved the project itself but also enhanced my skills as a researcher. This process of continuous reflection is essential in any academic or professional setting, fostering a culture of excellence and ongoing learning.