PHP2550 Final Reflection

Yiwen Liang

12/15/2023

This is the reflection report that aims to discuss the changes made to each project, the reasons behind these changes, and the lessons learned through the improvement process. The original and updated reports and code files for each project are located in the corresponding folder in the GitHub portfolio: https://github.com/yiwen-liang/PHP_2550_Final_Portfolio. Each folder has a separate README file that introduces the project in details, and each README file includes an abstract of the report, the guidelines of the documents available in this folder, data availability, acknowledgement, and environment version information that has all the packages utilized in this project and their versions presented.

In this report, the focus will be on highlighting specific changes made to each project. Common modifications, such as addressing grammar mistakes and typos and adjusting text descriptions to align with changes in content and format, will not be reiterated in each project section below.

Project 1

The first project focuses on conducting an exploratory data analysis to investigate the impact of smoking during pregnancy (SDP) and exposure to environmental tobacco smoke (ETS) on adolescent self-regulation, externalizing behavior, and substance use. Given that this marks the first project in this course, there is room for enhancement in various aspects, encompassing format, language, and analysis. The revised report is now structured in a conventional journal format, comprising sections such as Abstract, Introduction, Methods, Results, Discussions, and Conclusions. A more formal title has been crafted, and meticulous attention has been given to generating appropriate references to journal articles.

In an exploratory data analysis, incorporating additional tables and figures is crucial to substantiate statements and draw preliminary conclusions. With 77 variables in this dataset, presenting information about every variable in a single report is impractical and unnecessary. Therefore, another type of modification made to this project is related to the selection and presentation of tables and figures. For instance, I created a table that includes summary statistics for more than 40 variables in the original report. In this table, all the information is stacked together, making it challenging for the reader to grasp the details at first glance,

and it becomes unclear which variables are essential for our analysis. After modification, I provide a verbal description of the summary statistics for key demographic variables, while only displaying the summary of variables pertaining to self-regulation and externalizing issues in a table. An additional example is the table illustrating the summary of responses related to child outcomes by SDP and ETS at different time points. I generated one table for each period, leading to a total of 7 plots with similar formats and related content. Now, I have consolidated comparisons by prenatal and postnatal exposures into two composite tables. This approach allows readers to clearly identify the specific exposure they are examining and facilitates the comparison of the average responses of children stratified by mothers' smoking status at different times during pregnancy.

I also make a minor addition by including the interrelations between prenatal and postnatal smoke exposure, as well as interrelations among variables related to self-regulation issues. This was missed in the initial report, and I have now addressed this for exploration in the updated version, as suggested by Dr. Lauren Micalizzi.

In the final aspect of the improvement, I introduced summary variables for both prenatal and postnatal smoke exposure, denoted as SDP and ETS, along with their corresponding intensities $SDP\_intensity$ and $ETS\_intensity$. The addition of these newly-created variables allows for a comprehensive characterization of overall SDP and ETS, thereby preventing the need for repetitive and redundant comparison plots and tables. In the initial report, I included over 50 small bar plots depicting various types of substances, the smoking status of mothers during trimesters of pregnancy, and children's exposure to environmental tobacco smoke at an early age.

The feedback from the first project has been instrumental in guiding us on how to present our findings more clearly, systematically, and professionally. This constructive feedback not only shaped the subsequent two projects but also provided a valuable framework for tasks in future work and learning. In the initial stages of scientific research, gaining knowledge about the types and common analytical perspectives of exploratory data analysis (EDA) enables us to develop a preliminary understanding of the entire dataset. This systematic analysis of data is particularly valuable when tackling a new research problem.

Project 2

The second project aims to develop a regression model for predicting the composite outcome of tracheotomy and death in neonates with severe bronchopulmonary dysplasia (sBPD). Incorporating feedback from the first project, several format issues in Project 2 were addressed. Notably, modifications were made to include the Results and Conclusions in the Abstract. Additionally, a new Data section was introduced before the Methods section. The primary objective of this project is to create a prediction model, and the selection of the

regression model and variables is heavily influenced by the dataset's characteristics, including both predictors and outcomes. To commence the process, a concise exploratory data analysis (EDA) was conducted, and the data pre-processing steps are detailed in the report.

The primary modification made to this project involves the development of predictive models. In the initial report, the crucial element of including the center as a random effect was omitted. Given that the data is sourced from multiple sites, considering it as multilevel data and fitting mixed-effect models is essential. Consequently, two prediction models for the composite outcome of tracheotomy and death in infants with severe bronchopulmonary dysplasia (sBPD) at 36 weeks and 44 weeks were constructed. For both the week-36 and week-44 models, the approach involved using lasso and the best subset for variable selection of fixed effects, followed by fitting a mixed-effects model using the selected fixed effects and incorporating a random intercept for each center. In the initial report, I considered lasso, ridge, and best subset, three approaches introduced in class for variable selection. However, in the updated report, I excluded ridge regression. This decision was driven by the fact that ridge regression does not shrink the coefficient estimates toward zero, making it unsuitable for obtaining a subset of "preferred" variables, which slightly deviates from our primary goal and the planned model construction.

There are still limitations to this report after modification. Another commonly used approach among classmates is the 'glmmlasso' package. However, due to limited time, I did not manage to figure out a proper way to perform cross-validation using 'glmmlasso'. Regarding the process I opted for, which first conducts variable selection of fixed effects and then fits a mixed-effects model using the selected fixed effects, some scientists argue that separating the fixed and random effects while conducting variable selection may be problematic, as the structure of the random effects will have an influence on which fixed effect variables are selected. Personally, I consider this project assignment to be the most challenging among the three, and there's certainly room for improvement in the pre-processing of the data, techniques for variable selection, and the structure of the models.

Project 3

The third project is centered on conducting simulation studies to explore the applicability of the transportability analysis of the prediction model to simulated data in the absence of full data. Through evaluating the performance of the cardiovascular disease (CVD) prediction model developed from the population in the Framingham Heart Study, after transporting to the population from the National Health and Nutrition Examination Survey (NHANES) in 2017, and to the simulated target population, we plan to gain valuable insights into the generalizability of this model and hence answer the research question. This project is more straightforward compared to the previous ones, and as it is the final project, fewer changes are made.

I will commence by revisiting the aim of this project. As previously outlined, the objective is to use the CVD model as a case study to explore the applicability of transportability analysis to simulated data when full data is not accessible. During the coding process of the Brier estimator and the simulations, I found myself gradually deviating from the initial focus of the study. This serves as a reminder of the importance of establishing clear research objectives before initiating a study and maintaining focus throughout the research process.

Another significant modification pertains to the data-generating mechanism in simulation studies. Initially, the approach involved simulating each variable independently using its own summary statistics and distribution in the source population, assuming independence among variables. However, recognizing that assuming independence among health-related variables results in a loss of information and an inaccurate simulation of the target population, adjustments were made to better capture the interrelations among variables. Hence, I introduced a second attempt in the report, incorporating the use of a multivariate normal distribution to retain correlations between variables. Notably, continuous variables initially displayed right-skewed distributions, which were addressed through log-transformation to achieve approximately normal distributions. The simulation for categorical variables also considered associations with both categorical and continuous variables, drawing from the multivariate normal distribution to preserve these connections. Subsequently, the continuous results were converted back to binary categories based on quantiles for the three binary variables. It turns out that the estimated Brier score from the simulation adopting the multivariate normal distribution is the closest to the Brier score estimates obtained from the original NHANES data. Maintaining these correlations in the simulation results in a better representation of the target population.

Here are some other minor adjustments. Firstly, there were typos in the equation for the Brier estimator and the definition of the testing indicator, D, in the initial report. These elements are crucial for presenting the key estimator in this analysis, and I will ensure greater accuracy in their representation in this report and in future work. Additionally, regarding the layout of summary tables for the source population (Framingham) and the target population (NHANES), I initially created two separate tables with slightly different sets of variables. This made it challenging for readers to compare the two datasets effectively. Recognizing the importance of summary statistics for NHANES and the need to reference the distribution and covariance matrix of the same set of variables in the Framingham data for simulation, I have now combined the two tables side by side in the updated report. This modification enhances clarity and underscores why transportability analysis of the prediction model is both necessary and beneficial.

Monte Carlo simulation is a widely employed technique across various fields, yet it's notable that we have only covered one worksheet specifically dedicated to this topic. The third project is appreciated for its exclusive focus on simulation studies, offering an additional opportunity to apply simulation techniques, particularly the ADEMP framework, and design simulations tailored to a real-life dataset. Upon revisiting

and modifying this project, and the two projects ahead, I've observed enhancements in my ability to conduct statistical analysis, compose a scientific report, and, most importantly, think analytically and professionally.