Making the Most of Large Datasets: Meaningful Exploratory Analysis in Observational Studies

Matthew Spivey (W’ 27), under the leadership of Dylan Small (W, Professor, Statistics and Data Science) and Will Bekerman (W, PhD Student, Statistics and Data Science), funded by the Penn Undergraduate Research Mentoring Program

Motivation

We are analyzing the Wisconsin Longitudinal Study, which was conducted on Wisconsin high school graduates of the year 1957 across multiple decades. Our motivating example is to measure how parental alcoholism impacts children later in life. We are focusing on five categories of outcomes: Economic Outcomes, Physical Health, Mental Health, Substance Use, and Interpersonal Relationships. These categories will help us systematically assess how parental alcoholism influences financial stability, health, psychological well-being, substance use, and relationships, allowing for a targeted and insightful analysis.

Method

To ensure the robustness and replicability of our findings, we will divide the dataset into two independent subsets. We will analyze each subset separately and conduct exploratory data analysis while maintaining a familywise error rate of 0.025. This approach was designed to minimize bias and increase the reliability of the results. By comparing and synthesizing the outcomes, similar to a two-teamed approach, we can achieve a more comprehensive understanding of how parental alcoholism affects various life outcomes.

Process

We are re-evaluating the data selected for the exploratory data analysis, writing the protocol for the analysis, and confirming our method of splitting the data. As planned, we are on track to finish this project by the end of the fall semester.

Once our work is completed, the results will not only provide information about our motivating example but also confirm the process of splitting data to perform exploratory data analysis.

PURM
Penn Undergraduate Research
Mentoring Program