Michelle Casas

Stats 381: Statistics for open science

November 2, 2024

A yellow and purple squares

Description automatically generated

For this assignment, we were asked to replicate [Wesley Jeffrey’s figure](https://journals.sagepub.com/doi/10.1177/23780231241271717) on intergenerational educational mobility across five birth cohorts. Jeffrey did not include a replication package for this analysis but included a related supplemental document with two paragraphs giving a broad overview of the data used and analysis. Despite this document, recreating this figure was extremely hard for many reasons. First, because Jeffrey did not include a replication package, all of the data had to be downloaded from GSS. This meant creating an account with GSS and combing through all the variables to find the ones he used.

Second, new variables needed to be calculated from the original variables, which he does not indicate in his materials, but becomes clear as you start working with the data. Variables that are key to the analysis include the respondent’s birth year (in order to determine the decade born), decade (to group by in the analysis), and all of the statistic variables used in creating the figure including conditional probability and the correlation coefficient.

Third, it is unclear how Jeffrey handled missing variables, or how he cleaned his variables. For example, it is unclear if he included respondents between the ages of 30-40 (indicated by his supplemental information) or those > 30 (in his publication, he writes that he considered anyone born before 1990)?

Fourth, it was hard to get the plots to look like his. In the main plot, I was unable to get the decades to show within each pair of educational attainment for respondent and parent. My smaller plots along the bottom also don’t match his. From the publication, it sounds like the max equality and max inequality smaller plots were fake plots that were inserted to show the two extremes. The “overall” plot was a smaller version of the main plot reinserted for comparison purposes. While helpful to visualize, the smaller plots along the bottom were incredibly time consuming to produce.

Finally, my rho calculations were 0.4189, or approximately 0.42, whereas Jeffrey’s rho is 0.43. These values are not too far off, but I may have calculated the correlation coefficient incorrectly and it would have been helpful to have Jeffrey’s code to see what they did.

Overall, the hardest part of this exercise was how time consuming it was because we had to do this analysis from scratch. Compared to the assignment where we had a replication package, this took significantly longer, and the final result still does not look like Jeffrey’s. This assignment underscores the importance of sharing analysis files for reproducibility; if we want to confirm Jeffrey’s, or anyone else’s, findings we would have a difficult time doing so without the data and code used.