An Exercise in Data Analysis and Statistical Inference

Miles D. Williams

May 20, 2019

In this exercise, I (1) offer my best guess for the average treatment effect of the intervention and (2) summarize the evidence against the claim that the intervention had no effect. For both 1 and 2 I follow as closely as possible the standard operating procedure (SOP) suggested by Don Greene's lab at Columbia University. As the database specialist who compiled the data at hand is not available, and I do not have access to the pre-registered analysis plan, turning to Don Greene's SOP offers a useful starting point for identifying appropriate estimators and methods for statistical inference.

Making Sense of the Data

The database specialist left few clues about this data before leaving. She left behind three datasets, one that contains baseline data on 152 individuals, one that contains information of the randomization design for the policy intervention, and one that contains endline data for two outcome variables (one continuous, `Yc`, and one binary, `Yb`). The baseline and endline data oddly have five duplicate rows at the bottom, which I removed prior to merging the data.

Identifying Variables

The baseline data contains 19 columns of covariates for each individual in the dataset. All are simply labeled with an `X` followed by a number (ranging from 1 to 19). All that is known initially is that `X1` captures an individual's "status," which could refer to marital status, though it could also mean any number of things such as urban-rural status. `X2` captures age, though `X3` and `X4` are nearly identical to `X2`. `X3` seems to reflect an updated data draw on individuals' ages, and `X4` a final, more complete vector of ages. `X9` and `X14` both denote income; though, not the same income. One thought is that one of the variables is a corrected vector of income data, but another plausible guess is that these reflect dual incomes for a single household. `X8` denotes groups, ranging from 'a' to 'e'. `X11` and `X19` denote variables called `FISMOCheck` and `StatoFix2001` respectively; though, no description is given for what these variables are. My best guess is that `FISMOCheck` refers to a flexible single-master operation check for domain roles and naming on Windows, and values (`D` followed by a numerical value) probably represent domains relevant for certain aspects data transfer when the specialist was collecting/organizing the baseline dataset. It's less clear to me what `StatoFix2001` is; however, my guess is that it may have something to do with joining domains between a Windows operating system and some other operating system. The identities of the remaining covariates are unknown.

Dealing with Missing Data

Many of the covariates in the baseline dataset have missing values. Following the SOP, as less than 10 percent of the values in each variable is missing, I impute missing values for each of the numeric covariates with the average value for that variable among cases with non-missing data. I depart from the SOP, however, by using block-specific averages for imputed values. I did this because I noticed a considerable degree of clustering per block in values of many of the covariates. As I anticipate missing values for observations are likely to track with the block average, I rely on this value rather than the total sample average as a more accurate guess at the expected value of missing data.

Though I address missingness in baseline data as described above, attrition (missingness) remains an issue for each of the outcome variables in the endline data. There is a roughly 7.9 percent attrition rate for `Yc`, the continuous outcome variable, and a roughly 8.6 percent attrition rate for `Yb`, the binary outcome variable. An initial concern is whether attrition rates are significantly different than predicted by chance between treatment and control arms of the policy intervention, or among blocks. Following the analysis to follow, I conduct a heteroskedasticity robust F-test, implemented with Studentized permutation, to assess whether attrition rates are asymmetrically distributed in the data. Details of that analysis can be found there. For now, I will only say that attrition does not appear to be significantly different between treatment and control groups, or among blocks.

Choice of Estimator and Adjustment Strategy

Statistical Inference

Analysis and Results