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 Green'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 Green's SOP offers a useful starting point for identifying appropriate estimators and methods for statistical inference.

1 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.

1.1 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.

1.2 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.

2 Choice of Estimator, Adjustment Strategy, and Specification of Outcomes

In considering choice of estimator and adjustment strategy, I refer to the SOP for guidance. The SOP relies on OLS regression as the default estimator for generating estimates for average treatment effects. This is a straightforward method of obtaining differences in means between treatment and control groups, so I rely on OLS as suggested.

In devising an adjustment strategy, I consider both the data at hand and the randomization design of the policy intervention. The trail randomized the intervention within blocks with different treatment probabilities per block. Blocks 1, 2, and 3 only had 4 individuals each, with 2 (half) in each block randomly assigned treatment. Blocks 4, 5, and 6 similarly had half of the individuals assigned to treatment; though 8 individuals in total were in each. Blocks 7 to 9 had 12 individuals each, but only 1/3 of the subjects in each were given treatment. Finally, block 10, by far the largest, had 80 total individuals, only 5 percent of whom received treatment. The SOP advises two strategies for such a design. The first is to use OLS to regress the outcome on (1) a treatment indicator, (2) a set of block indicator variables with one block dropped to serve as a reference category, and (3) a set of treatment block interactions where the proportion of observations within each block is subtracted from its respective 0-1 block indicator. 3 is the equivalent of mean centering the block indicators.

The second approach is to estimate a least squares dummy variable (LSDV) regression—essentially an OLS model where the outcome is regressed on the treatment indicator and block indicators (minus a reference category) without interaction. The SOP recommends this approach be used under the extreme condition that, for at least one block j, the following inequality holds:

$$\frac{N_j}{\sum_j N_j} > 20 \cdot \frac{N_j P_j (1 - P_j)}{\sum_j N_j P_j (1 - P_j)}.$$

In the above, N_j denotes the number of subjects per block j and P_j the probability of treatment per block j.

To determine which of these two approaches to use, I calculate the left and right sides of the above inequality for each of the 10 blocks included in the data. I find that in 2 blocks, this inequality is met. I therefore rely on the second approach, estimating the treatment effect without treatment-block interactions.

I further have several baseline covariates per observation. Though in theory randomization within blocks should be independent of both observed and unobserved unit characteristics, adjusting for covariates in estimating the average treatment effect is justified on the basis that including covariates that are strongly correlated with the outcome, regardless of their association with treatment assignment, help to improve statistical power. The SOP offers guidelines for how to adjust for covariates in the analysis, as well as the number of covariates to include.

The first thing to consider is the number of individuals assigned to treatment. Let M denote individuals assigned to the treatment arm. If $M \geq 20$, the SOP recommends adjusting for covariates by regressing the outcome on (1) the treatment indicator, (2) the covariates, and (3) interactions between the treatment indicator and mean-centered values of the covariates.

If $M < 20 \leq N$, where N is the total number of observations, the SOP recommends regressing the outcome on items 1 and 2 described above, and not including treatment covariate interactions.

Finally, in the extreme case where $N < 20$, the SOP recommends estimating only the difference in means between control and treatment groups.

After attrition rates are accounted for, $N = 140$ for the continuous outcome, with $M = 32$, while $N = 139$ with $M = 32$ for the binary outcome. The SOP therefore calls for the first adjustment strategy, that is, regressing the outcome on the treatment indicator, covariates, and treatment-covariate interactions (with covariates mean-centered).

Though in an experiment intuition seems to favor a less complicated adjustment strategy (why not just estimate the average treatment effect?), as @lin2012a describes this specification proves more robust to bias induced due to improper specification of the association between covariates and the outcome. OLS assumes a linear association between predictors and outcome, but certain covariates may have other sorts of associations with the outcome (e.g., quadratic, log-linear, log-log, etc.). Estimating the average treatment effect conditional on holding covariates at their mean, as @lin2012a shows, proves surprisingly effective at mitigating, at the very least, bias induced by improper specification. I therefore follow this strategy in analyzing the data at hand.

Finally, the number and choice of covariates to adjust for needs to be justified. The SOP recommends including no more than $M/20$ covariates when using interactions and no more than $N/20$ when not using interactions. As I have decided to go with the interaction adjustment strategy, I restrict the number of covariates to no more than $M/20 = 1.6$. Since I can't include 6/10 of a covariate, I include only 1. This procedure for limiting covariates, in addition to helping to simplifying the analysis, helps to steer practitionors clear of the temptation to use "kitchenskink" regressions.

Restricting the number of covariates to 1 means I must be extra choosy in identifying the best covariate to include. @bruhnMcKenzie09 suggest a covariate that is highly correlated with the outcome of interest, no matter its distribution between treatment and control arms of the study. Some very basic bivariate correlations show that `X4`, subject age, is highly correlated with both the continuous and binary outcomes. Pearson's ρ equals 0.51 for age and the continuous outcome and -0.52 for age and the binary outcome. These estimated ρs are much larger relative to any of the other covariates included in the data.

Given the above discussion, I therefore generate estimates of the average treatment effect of the intervention on the outcomes of interest using the following specifications to be estimated by OLS:

$$\ln(Y_i^c) = \beta_1 z_i + \beta_2 x_i + \beta_3 z_i \cdot (x_i - \bar{x}_j) + \mathbf{B}\alpha + \varepsilon_i, \tag{1}$$

$$Y_i^b = \gamma_1 z_i + \gamma_2 x_i + \gamma_3 z_i \cdot (x_i - \bar{x}_j) + \mathbf{B}\eta + \upsilon_i, \tag{2}$$

where β_1 and γ_1 denote the ATE for the block-randomized policy intervention z, denoted by `Zdesign` in the data, on the continuous and binary outcome variables respectively. x_i is subject age in years and $\bar{x}_j$ is the average age of subjects per block j. I mean-center age on the block-specific average rather than the total average so that estimates reflect the ATE per block. $\mathbf{B}$ is a vector of block indicators where I drop block $j = 1$ to serve as the reference category.

Though not recommended by the SOP, I log-transform Y_i^c in equation 1. The continuous outcome variable has a highly skewed distribution. Values range from 1.24 to 67.16, though the mean is 7.28 and the median is 3.56. This skewness poses challenges for straightforwardly estimating the average treatment effect. Log-transforming values helps to overcome this issue without loss of data or the need to resort to methods such as robust regression. This choice does slightly change the interpretation of the estimand for the treatment affect, however. As this model has the functional form of a log-linear model, the estimated parameter on the treatment variable can be interpreted as the percent change in the outcome given treatment. More precisely, $\%\Delta Y_i^c = 100 \cdot (e^{\beta_1} - 1)$.

An alternative to using the natural log of the continuous outcome is a rank-transformation of values. Though a viable choice, the drawback of a rank-based transformation is that the ATE will no longer have a straightforward, substantive, interpretation. With log-transformation I can still generate an ATE with a near direct translation into the value of the outcome under treatment versus control. With rank-transformation, the value or magnitude of the ATE looses its meaning.

3 Statistical Inference

4 Results

5 Checks for Infelicities in Design