An exercise in data analysis and statistical inference

Jake Bowers, Nuole Chen, Russell Burnett

17 May, 2019

We commonly ask two questions about a given policy intervention: “What is our best guess about the average treatment effect of the intervention?” and “What is the evidence against the claim that our intervention had no effects?”

Please provide your one best answer to each of those two questions for the simulated intervention in the data provided along with this exercise. We are particularly interested in how you think about your choice of estimators for the underlying causal effect(s) and your choice of tests/test statistics used for confidence intervals and/or hypothesis tests. The OES justifies its choice of statistical procedures based on the design of the randomized experiments that we analyze, so we encourage you to use a design-based approach in your analysis.[1] We are also interested to read how you justify the decisions you make about the data itself. Please keep your answer to 5 single-spaced pages.

To give you a taste of the OES Methods Team reanalysis work, we have simulated a dataset that represents some of the common challenges we face when using administrative data to produce statistical inferences about the causal effects of a randomized policy intervention.

You can learn a lot about the study design from the data themselves:

- `design.csv` contains the subject `id` number; the indicator for the block or stratum `b` within which the new intervention was randomized; and `Zdesign`, which indicates randomized assignment to the new intervention (`Zdesign=1`) or the status quo (`Zdesign=0`). You will notice that the treatment was expensive: relatively few units were assigned to the treatment condition in each block, regardless of the size of the block.
- `outcomes.csv` contains the subject `id` number and two outcomes measured post-treatment (`Yc` and `Yb`).
- `baseline.csv` contains the subject `id` number and a set of covariates measured before the experiment was fielded (variables that all start with `X`).

The OES Methods Team and primary analysts commonly work with data of this form — the baseline data may have come from a database pull, the design data (random assignment and blocks) may be generated by the OES team, and the outcomes are shipped back to the OES from the agency.

We often confront confusing and incomplete data documentation. For this exercise the baseline covariates are very sparsely documented. We don’t know why there are two Income variables or what FISMOCheck and StatoFix2001 refer to or why certain covariates appear to be near copies of each other. In this exercise, imagine that the database specialist in the relevant agency has left the federal government and nobody knows what she meant by those labels or how she coded the variables in the covariate set. Here is what we know:

- X1: Status
- X2: Age
- X9: Income
- X14: Income
- X8: Group
- X11: FISMOCheck
- X19: StatoFix2001

You will notice that these data have missing data and other infelicities — such data problems are a common part of OES life. For example, you might run into problems merging the datasets together.

[1] On this see for example Chapter 3 of Gerber and Green (2012) and https://github.com/acoppock/Green-Lab-SOP.

References

Gerber, Alan S, and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation.* WW Norton.