HW 1: Exploring Causal Inference in Experimental and Observational Studies

See the course syllabus for more instructions about working in teams. Students should turn in individual write-ups but may collaborate on code.

DUE: Sunday May 6, 11:59 pm, on canvas.

Getting Data and HW Overview

For this assignment, you should start with data from a randomized experiment. You are welcome to work with the dataset provided for the warm-up exercise, but several additional datasets are provided on github. You may also bring your own dataset from another source. You can change datasets from assignment to assignment; you will be asked to complete this first assignment about average treatment effects, and then a second assignment about heterogeneous treatment effects, and it may be easiest to use the same dataset for both.

For the provided datasets, you should first take a quick look at the associated papers to understand the data. Also take a look at the key results. You will end up facing different types of issues depending on data set size, number of covariates, and strength of the basic treatment effect results. For example, in the paper on Charitable Giving, using ML methods to look for treatment effect heterogeneity in our experience often lead to spurious results, as the baseline results are not that strong. So if you pick this dataset, you may be emphasizing lack of findings or false findings rather than positive findings. On the other hand, some of the other datasets have richer or stronger results and heterogeneity, and/or more covariates. A few have some additional complications. (E.g. the paper on social voting has multiple observations per household; we suggest for simplicity analyzing only one household member so you don't have to worry about independence of observations.) Many have multiple outcomes and multiple treatments, so you'll want to pick one of each to focus on. It is fine to use linear models for binary outcomes for the purposes of this class, or you can also use logistic versions of procedures if you like.

You should not assume that because a dataset is included, it has a particular type of results associated with it. There aren't that many large-scale, publicly available randomized experiments out there. (If you find more that look interesting, please send us an email with a link or source, as we'd love to build up our collection. We intentionally didn't include the Lalonde data here because it is over-studied and the experiment is small.)

Your assignment is to test out some different methods for estimating average treatment effects. To do so, the first part of the assignment is to turn the randomized experiment into an observational study. I'd like you to systematically delete some observations as a function of X's and treatment status (but not as a function of observed outcomes), and give a little thought to how you would like to do so in order to make things interesting. In general, X's that are correlated with both treatment assignment and outcomes create challenges for causal inference. This part of the exercise is a good warm-up to thinking about causal inference, as it helps you think about how different data-generating processes lead to observational datasets and associated biases in estimation.

In addition to the code in the warm-up exercise, you can also run our tutorial on ATEs, which provides example use of many of the methods.

Specific Assignment

For your assignment:

1. Before beginning, estimate the average treatment effect and the confidence interval in the randomized experiment.
2. Describe your method(s) for systematically deleting some observations as a function of X's and treatment status (but not as a function of observed outcomes), provide summary statistics, and show the old and new (simple average estimated) treatment effect.
    - Try to drop enough observations (and with an aggressive enough rule) that your new point estimate of the treatment effect is significantly different than the point estimate in the full randomized experiment. This is not a hard and fast rule, just a guideline.
    - Also make sure your rule is complex enough that it is not trivial to recover the propensity score; for example include some nonlinearities and multiple variables.
    - If all the methods below give the same, correct answer, try a more aggressive or complex dropping rule.
    - Be sure to preserve *overlap*, so that there are no X values from which you can deterministically read off that W = 0 or W = 1. However, deleting some observations such as to make propensities *close* to 0 or 1 is a great way to stretch methods and test their robustness.
    - Plot the histogram of the bias function as in Athey, Imbens, Pham and Wager (AER P&P, 2017) to summarize how challenging your problem is after dropping observations
3. In the modified dataset, estimate the ATE (note that most of the code required to do the exercises below can be written by combining bits of code from the warm-up R exercise, and the tutorial):
    - Test out the following traditional methods for estimating the ATE: (i) propensity score weighting via logistic regression; (ii) direct regression analysis via OLS; (iii) traditional double robust analysis via augmented inverse-propensity score weighting that combines the above two estimators.
    - Now, make the dataset more high dimensional. You could add interactions (as in the tutorial), non-linearities, or both. Estimate the outcome and propensity models using the lasso, and try the 3 approaches above again. With the lasso, choice of tuning parameter is important.
        - What happens if you cross-validate for lambda?
        - Try training the outcome and propensity models for a wide grid of lambda (you should start from one extreme where lambda is so small that the model gets unstable, and then make lambda bigger until all coefficients are 0). How do the 3 estimators you computed change with lambda? In order to just do just 1 grid search, you can simultaneously multiply the cross-validated choice of lambda by the same number for each lasso you run (i.e., for outcomes, propensities).
    - Repeat the above steps again, but now using the machine learning method of your choice (you can choose from trees, forests, boosting, kernel regression, deep nets, etc.) Does cross-fitting make a difference in this case?
        - Try varying the sample size n available to the method (e.g., by deleting various fractions of the training data at random). How does performance change?

4. Next, consider estimation of the average treatment effect on the treated (ATT)
    - Estimate the ATT via the direct OLS-based method. Compare your estimate to the ATE estimate obtained in part 3.
    - Estimate the ATT via forests, as discussed in class.
    - Implement a variant of the Belloni-Chenozhukov-Hansen double selection method for the ATT, on the high-dimensional version of your problem for part 3(b).
5. Compare and interpret your results.

Your write-up should include code with output, as well as an electronic document (submitted individually) that discusses the results. The code output should preferably be in a knit document (if it works, generated by "knitting" as per the R instructions, but this sometimes fails, it is finicky), or else attached separately as plots and tables. Try to make your document self-contained by adding in figures and referring to specific numbers/standard errors in the text where relevant. If you worked with group members on your code, indicate the group members on the assignment, but your write-ups should be done individually, and each member should submit the code/knit file.