Info 6751. Fall 2022. Problem Set 5. Due on Canvas by 5pm on 3 Oct.

This problem set is about **matching**. It is different from other problem sets in two ways.

- There is only one part. We are covering this topic on both Tuesday and on Thursday. When you are stuck, the answer is likely in Stuart (2010).
- This problem set is very self-guided. There are many matching methods, and you are asked to choose one method and run with it. We will all get different answers! That's ok.

Data and Causal Assumptions

This problem set uses data from Dehejia & Wahba (1999, 2002), a version of which we have used previously. For this problem set, download the data `nsw.dta` from http://www.nber.org/~rdehejia/data/nsw_dw.dta. These data contain 260 untreated individuals and 185 individuals who received a job training intervention. The data are described here. For the problem set, here is an abbreviated description:

The outcome variable is `re78` (earnings in 1978).
The treatment variable is `treat`: (1 job training, 0 none).
Pre-treatment covariates include:

- `age`: numeric
- `education`: numeric, number of years
- `nodegree`: 1 if no high school degree, 0 otherwise. This is a dichotomized version of `education`
- `black`: 1 if Black, 0 otherwise
- `hispanic`: 1 if Hispanic, 0 otherwise
- `married`: 1 if married, 0 otherwise
- `re74`: earnings in 1974
- `re75`: earnings in 1975

The data also include a constant `data_id` which simply identifies the dataset.

Causal Assumptions and Estimand

Our goal is to estimate the Feasible Sample Average Treatment Effect on the Treated (FSATT),

$$\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\left(Y_i^1 - Y_i^0\right)$$

where $\mathcal{S}$ is the set of matched treated units. In this set, use the observed values for Y_i^1 and use the average of matched controls to impute Y_i^0. This set may not include all treated units, because some may be dropped in question (2).

Throughout, assume consistency, exchangeability, and positivity given these pre-treatment covariates.

Rubric for Grading

Everyone will have different answers. There are two important requirements that will affect grading:

- Explain as though writing to an undergrad who has taken one semester of introductory statistics.
- Include your code, either embedded in the PDF (e.g., with RMarkdown) or as a separate uploaded file.

Feel free to use one of the many software implementations for matching. You can also code from scratch.

Questions

1. (10 points) Define a distance metric for matching.
    - How do you measure the distance between two distinct covariate vectors $\vec{\ell}$ and $\vec{\ell}'$?
    - Define in math and in words. Motivate your choice.
    - Examples include Euclidean distance, Manhattan distance, Mahalanobis distance, squared difference in propensity scores, etc. For coarsened exact matching, this would be a distance of 0 if in the same coarsened stratum and ∞ if not.
2. (10 points) Define a caliper.
    - At what distance between units i and j would you say that the two are too far apart to consider as matches?
    - How many treated units do you lose by applying the caliper?
3. (10 points) Implement a matching method.
    - How many control units are matched to each treated unit?
        - e.g. 1:1, 2:1, a varying number depending on the pool, etc.
    - Are you matching with or without replacement?
    - Is your matching algorithm greedy or optimal?
4. (10 points) Evaluate the matched sample.
    - How many treated units do you have?
    - How many control units?
    - How similar are the covariate distributions of the treated and control units in the matched sample? Ideally you would do this for all covariates, but to restrict the length of the homework summarize the mean value of `education` in the matched sample by treatment category.
5. (10 points) Analyze the outcome. What is your FSATT estimate?