Program Evaluation (Causal Inference) 2: Matching

Instructor: Yuta Toyama

Last updated: 2020-06-22

Section 1

Introduction

Introduction: Matching Estimator

- Idea: Compare **individuals with the same characteristics** X across treatment and control groups
- Key assumption: Treatment is random once we control for the observed characteristics.
- Do you remember we already learnt a similar idea before?

Section 2

Identification

Matching

- Let X_i denote the observed characteristics:
    - age, income, education, race, etc..
- Assumption 1:

$$D_i \perp (Y_{0i}, Y_{1i}) | X_i$$

    - Conditional on X_i, no selection bias.
    - Selection on observables assumption / ignorability
- Assumption 2: Overlap assumption

$$P(D_i = 1 | X_i = x) \in (0, 1) \ \forall x$$

    - Given x, we should be able to observe people from both control and treatment group.
    - We call $P(D_i = 1 | X_i = x)$ **propensity score**.

Identification

- The assumption implies that

$$E[Y_{1i}|D_i = 1, X_i] = E[Y_{1i}|D_i = 0, X_i] = E[Y_{1i}|X_i]$$
$$E[Y_{0i}|D_i = 1, X_i] = E[Y_{0i}|D_i = 0, X_i] = E[Y_{0i}|X_i]$$

- The ATT for $X_i = x$ is given by

$$\begin{aligned} E[Y_{1i} - Y_{0i}|D_i = 1, X_i] &= E[Y_{1i}|D_i = 1, X_i] - E[Y_{0i}|D_i = 1, X_i] \\ &= E[Y_i|D_i = 1, X_i] - E[Y_{0i}|D_i = 0, X_i] \\ &= \underbrace{E[Y_i|D_i = 1, X_i]}_{avg\ with\ X_i\ in\ treatment} - \underbrace{E[Y_i|D_i = 0, X_i]}_{avg\ with\ X_i\ in\ control} \end{aligned}$$

- The components in the last line are identified (can be estimated).
- Intuition: Comparing the outcome across control and treatment groups after conditioning on X_i

ATT and ATE

- ATT is given by

$$\begin{aligned} ATT &= E[Y_{1i} - Y_{0i} | D_i = 1] \\ &= \int E[Y_{1i} - Y_{0i} | D_i = 1, X_i = x] f_{X_i}(x | D_i = 1) dx \\ &= E[Y_i | D_i = 1] - \int \left(E[Y_i | D_i = 0, X_i = x] \right) f_{X_i}(x | D_i = 1) \end{aligned}$$

- ATE is

$$\begin{aligned} ATE =& E[Y_{1i} - Y_{0i}] \\ &= \int E[Y_{1i} - Y_{0i} | X_i = x] f_{X_i}(x) dx \\ &= \int E[Y_i | D_i = 1, X_i = x] f_{X_i}(x) dx \\ &= + \int E[Y_i | D_i = 0, X_i = x] f_{X_i}(x) dx \end{aligned}$$

Section 3

Estimation

Estimation Methods

- We need to estimate $E[Y_i|D_i = 1, X_i = x]$ and $E[Y_i|D_i = 0, X_i = x]$
- Several ways to implement the above idea
  1. Regression: Nonparametric and Parametric
  2. Nearest neighborhood matching
  3. Propensity Score Matching

Approach 1: Regression, or Analogue Approach

- Let $\hat{\mu}_k(x)$ be an estimator of $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$ for $k \in \{0, 1\}$
- The analog estimators are

$$\hat{ATE} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)$$

$$\hat{ATT} = \frac{N^{-1} \sum_{i=1}^{N} D_i (Y_i - \hat{\mu}_0(X_i))}{N^{-1} \sum_{i=1}^{N} D_i}$$

- How to estimate $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$?

Nonparametric Estimation

- Suppose that $X_i \in \{x_1, \cdots, x_K\}$ is discrete with small K
  - Ex: two demographic characteristics (male/female, white/non-white). $K = 4$
- Then, a nonparametric binning estimator is

$$\hat{\mu}_k(x) = \frac{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\} Y_i}{\sum_{i=1}^{N} \mathbf{1}\{D_i = k, X_i = x\}}$$

- Here, I do not put any parametric assumption on $\mu_k(x) = E[Y_i | D_i = k, X_i = x]$.

Curse of dimensionality

- Issue: Poor performance if K is large due to many covariates.
- So many potential groups, too few observations for each group.
- With K variables, each of which takes L values, L^K possible groups (bins) in total.
- This is known as **curse of dimensionality**.
- Relatedly, if X is a continuous random variable, can use kernel regression.

Parametric Estimation, or going back to linear regression

- If you put parametric assumption such as

$$E[Y_i|D_i = 0, X_i = x] = \beta' x_i$$
$$E[Y_i|D_i = 1, X_i = x] = \beta' x_i + \tau_0$$

then, you will have a model

$$y_i = \beta' x_i + \tau D_i + \epsilon_i$$

- You can think the matching estimator as controlling for omitted variable bias by adding (many) covariates (control variables) x_i.
- This is one reason why matching estimator may not be preferred in empirical research.
  - Remember: Controlling for those covariates is of course important. This can be combined with other empirical strategies (IV, DID, etc).

Approach 2: $M-$Nearest Neighborhood Matching

- Idea: Find the counterpart in other group that is close to me.
- Define $\hat{y}_i(0)$ and $\hat{y}_i(1)$ be the estimator for (hypothetical) outcomes when treated and not treated.

$$\hat{y}_i(0) = \begin{cases} y_i & if \;\; D_i = 0 \\ \frac{1}{M} \sum_{j \in L_M(i)} y_j & if \;\; D_i = 1 \end{cases}$$

- $L_M(i)$ is the set of M individuals in the opposite group who are "close" to individual i
  - Several ways to define the distance between X_i and X_j, such as

$$dist(X_i, X_j) = ||X_i - X_j||^2$$

- Need to choose (1) M and (2) the measure of distance
  - R has several packages for this.

Approach 3: Propensity Score Matching

- Use propensity score $P(D_i = 1|X_i = x)$ as a distance to define who is the closest to me.
- Implementation:
  1. Estimate propensity score function by logit or probit using a flexible function of X_i.
  2. Calculate the propensity score for each observation. Use it to define the pair.