

# Causal Inference (6.S059/15.Co8/17.Co8)

Recitation, Week 5.

**Topic: Weighting and Regression**

Benjamín Muñoz

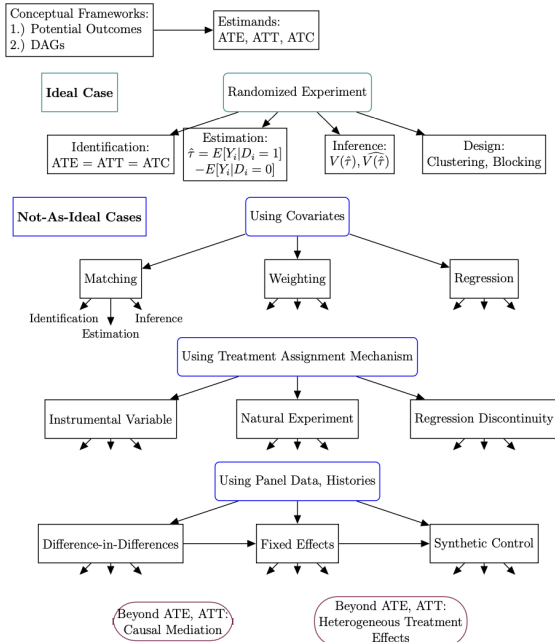
March 15, 2024

MIT

# Table of contents

1. Introduction
2. Selection on Observables
3. Weighting
4. Regression

# 1/ Introduction



# Identification and Estimation

- What if we observe a non-randomized treatment?
  - Risk of **confounding**:  $D_i \not\perp \{Y_{0i}, Y_{1i}\}$

$$\underbrace{\frac{1}{N_T} \sum_{i=1}^n (y_i | d_i = 1) - \frac{1}{N_C} \sum_{i=1}^n (y_i | d_i = 0)}_{\text{Observed Difference in Means}} = \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]}_{ATT} + \underbrace{\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]}_{\text{Selection Bias}}$$

- **Selection bias**: how different the treated and control groups are in terms of their potential outcome under control.
- ATT is **unidentified**: observed difference in means is a combination of two unknown quantities.

# Identification and Estimation

- Observational distribution  $\mathbb{P}$  of  $(Y_i, D_i, \mathbf{X}_i)$  and **counterfactual** (unobserved) distribution  $\mathbb{P}^*$  of  $(Y_{1i}, Y_{0i}, D_i, \mathbf{X}_i)$ .
- The causal quantities of interest ( $\Psi$ ) are functions of  $\mathbb{P}^*$ , but we get data from  $\mathbb{P}$ .
- A causal quantity  $\Psi$  is identified if we can write it as function of  $\mathbb{P}$ .

## 1. Identification

- A quantity is identified if it can be calculated using observed information.
- It is theoretically possible to learn the true value of that parameter with an infinite number of observations (Matzkin 2007, sec. 3.1).
- Set of **assumptions** that we are willing to make about the relationship between the observable and unobservable distributions.
- Identification tells us **what** to estimate, not **how**.

# Identification and Estimation

## 2. Estimation

- The process of learning about underlying causal quantities of interest using data.
- Statistical procedures devoted to approximate (best guess) the true value of  $\Psi$  with sample data (limited number of observations).
- **How** to do it.

## 3. Inference

- Quantification of uncertainty related to estimated point estimates (we only observe a portion of reality: sampling variability, assignment variability, etc.)

## **2/** Selection on Observables

# Selection on Observables

- It is an identification strategy adequate for an observational setting (no manipulation/nonrandomized treatment).
- There is some set of **pre-treatment** covariates such that treatment assignment is random conditional on these covariates.
- Given this set of “correct” covariates, we can use statistical adjustment methods (regression, matching, or weightin) to make conditional independence hold.
- It cannot be verified with observed data (Manski 2007).

## Assumptions

1. **Conditional Ignorability** (No unmeasured confounding, unconfoundedness, ignorability, selection on observables, no omitted variables, exogeneity, conditional exchangeable, etc.): Conditional on some covariates,  $D_i$  is (effectively) randomly assigned.
2. **Common Support** (Positivity): Treatment and control are both possible at every value of  $X_i$ .

# Identification Assumptions

Quantity of Interest: Average Treatment Effect (ATE)

1. **Conditional Ignorability:**  $\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp D_i | X_i = x$  for any  $x \in \mathcal{X}$ .
2. **Common Support:**  $0 < Pr(D_i = 1 | X_i = x) < 1$  for any  $x \in \mathcal{X}$ .

Quantity of Interest: Average Treatment Effect on the Treated (ATT)

1. **Conditional Ignorability:**  $\{Y_{0i}\} \perp\!\!\!\perp D_i | X_i = x$  for any  $x \in \mathcal{X}$ .
2. **Common Support:**  $Pr(D_i = 1 | X_i = x) < 1$  for any  $x \in \mathcal{X}$ .

Quantity of Interest: Average Treatment Effect on the Untreated (ATU)

1. **Conditional Ignorability:**  $\{Y_{1i}\} \perp\!\!\!\perp D_i | X_i = x$  for any  $x \in \mathcal{X}$ .
2. **Common Support:**  $0 < Pr(D_i = 1 | X_i = x) = Pr(D_i = 0 | X_i = x) < 1$  for any  $x \in \mathcal{X}$ .

# Estimation Strategies

1. **Subclassification:** basic estimator (only discrete covariates).
2. **Matching:** nonparametric imputation estimator that can handle continuous covariates. Create comparable groups by pairing (and dropping) observations.
  - ↪ impute missing potential outcomes using observed outcomes of “closest” units, based on some distance metric.
3. **Weighting:** continuous version of matching (instead of pairing, assign weights to units so that, on average, distributions are comparable).
  - ↪ uses the entire dataset, but multiplies each observation according to some weighting function.
4. **Regression:** adjustment of covariates (controlling) with a specific regression function.
  - ↪ commonly involves parametric assumptions (functional form).
5. **Combined Methods**

## 3/ Weighting

# Propensity Score

- **Propensity Score (PS):** probability of receiving treatment given  $\mathbf{X}_i$

$$\pi(\mathbf{X}_i) \equiv P(D_i = 1 | \mathbf{X}_i)$$

- Under a selection on observables strategy, the P.S. has a **balancing property**:  $\mathbf{X}_i \perp\!\!\!\perp D_i | \pi(\mathbf{X}_i)$
- Therefore, we can re-express **Conditional Ignorability**:  $\{Y_{1i}, Y_{0i}\} \perp\!\!\!\perp D_i | \pi(\mathbf{X}_i)$

## Estimation Procedure

1. Estimate  $\pi(\mathbf{X}_i)$  with a model (Binary  $D_i \rightsquigarrow$  logistic regression).
2. Check the resulting balance (if necessary, re-estimate the model).
3. Implement estimation technique (matching, weighting, regression).

# Weighting

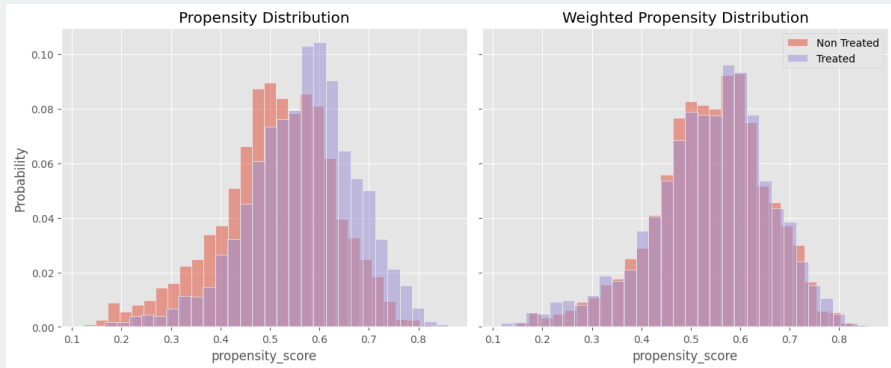
- Problems of matching: inefficiency ( $\Delta^- N$ ) and ineffective (crude tool to achieve balance).
- Matching can be expressed as a specific case of weighting:

$$\hat{\tau}_m = \frac{1}{n_1} \sum_{i=1}^n D_i \left( Y_i - \frac{1}{M} \sum_{j \in \mathcal{J}_M(i)} Y_j \right)$$

- Where  $\mathcal{J}(i)$  are the set of  $M$  closest control units to  $i$  in terms of  $\mathbf{X}_i$ .

$$\hat{\tau}_m = \frac{1}{n_1} \sum_{i:D=1} Y_i - \frac{1}{n_0} \sum_{i:D=0} \underbrace{\left( \frac{n_0}{n_1} \frac{K_M(i)}{M} \right)}_{Weight} Y_i$$

- Where  $K_M(i)$  is the number of times  $i$  is used as a match.



# Weighting Estimands

- Under **selection on observables**, we can identify two quantities of interest:

$$\star\tau_{ATE} = \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(\mathbf{X}_i)}{\pi(\mathbf{X}_i) \cdot [1 - \pi(\mathbf{X}_i)]}\right], \text{ where, if } D_i = 1, \text{ the weight is } \frac{1}{\pi(\mathbf{X}_i)},$$

and if  $D_i = 0$ , the weight is  $\frac{1}{1 - \pi(\mathbf{X}_i)}$ .

$$\star\tau_{ATT} = \frac{1}{P(D_i=1)} \cdot \mathbb{E}\left[Y_i \cdot \frac{D_i - \pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)}\right], \text{ where, if } D_i = 1, \text{ the weight is } 1,$$

and if  $D_i = 0$ , the weight is  $\frac{-\pi(\mathbf{X}_i)}{1 - \pi(\mathbf{X}_i)}$ .

# Weighting Estimators

- Horvitz-Thompson estimator: weight by **inverse propensity score** (inverse of the probability of being treated/untreated).

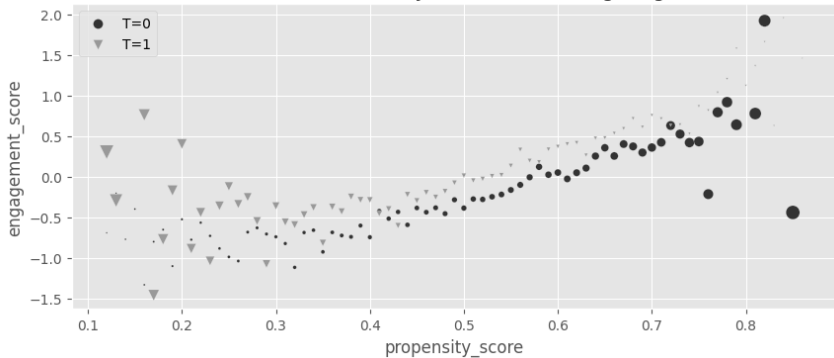
$$\hat{\tau}_{ATE} = \frac{1}{N} \sum_{i=1}^N \left[ Y_i \cdot \frac{D_i - \hat{\pi}(\mathbf{X}_i)}{\hat{\pi}(\mathbf{X}_i) \cdot [1 - \hat{\pi}(\mathbf{X}_i)]} \right] = \frac{1}{N} \sum_{i=1}^N \left[ \frac{D_i Y_i}{\hat{\pi}(\mathbf{X}_i)} - \frac{(1 - D_i) Y_i}{1 - \hat{\pi}(\mathbf{X}_i)} \right]$$
$$\hat{\tau}_{ATT} = \frac{1}{N_1} \sum_{i=1}^N \left[ Y_i \cdot \frac{D_i - \hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)} \right] = \frac{1}{N_1} \sum_{i=1}^N \left[ D_i Y_i - (1 - D_i) Y_i \frac{\hat{\pi}(\mathbf{X}_i)}{1 - \hat{\pi}(\mathbf{X}_i)} \right]$$

- **Logic:** upweight units with “rare” treatment values for their values of  $\mathbf{X}_i$ .
- Hajek estimator normalizes the denominator so the weights sum to 1.

# Practical Considerations

- IPW estimators are consistent but not unbiased for small  $N$ .
- **Check common support!**  $\rightsquigarrow$  large bias if  $\widehat{\pi}(\mathbf{X}_i)$  is close to the bounds (0 or 1). Also, high variance (unstable weights).
- Trimming procedures for extreme weights.
- Calculation of Variance? Bootstrapping or method of moments.
- Code Available: Google Colab.

Inverse Probability of Treatment Weighting



## 4/ Regression

# Regression

- Characterization of conditional probability distribution of  $Y$  for different levels of  $X$ .
- Typical focus: **Conditional Mean** or Conditional Expectation Function

$$\mathbb{E}[Y|X = x]$$

- How the average of  $Y$  varies across all possible levels of  $X$ .
- Non-parametric and parametric regression (one flavor: linear regression  $\leadsto$  OLS as estimator).

# Linear Regression in Randomized Experiments

1. **Estimation of Treatment Effect:** run linear regression (OLS estimator) for the outcome  $Y_i = \beta_0 + \beta_1 D_i + \epsilon$  (equivalent to Difference in Means).
  - ↪ Use robust standard error (HC2 type).
  - ↪ Randomization + Consistency = Linear Model.
2. **Balance Check:** run linear regression for the treatment  $D_i = \alpha_0 + \gamma X + \epsilon$ 
  - ↪ Goal: no statistically significant differences.
3. **Adjusting for Covariates:** run linear regression for the outcome  $Y_i = \alpha + \delta D_i + \gamma X + \epsilon$ 
  - ↪ Only pre-treatment covariates.
  - ↪ Only  $\delta$  has a causal interpretation.
  - ↪ ATE is consistent (small bias due to model misspecification = Partialling Out). Adequate control  $\Delta^+$  precision ( $\downarrow$  SEs).

# Linear Regression in Observational Studies

- Controlling for observed confounders.
- Pre-treatment covariates + adequate specification.
- Constant Effect and Functional form assumptions (**Linearity** and Additivity in OLS).
- Extra assumptions for interpretation of Causal Quantity (conditional-variance-weighted).