

Applied Methods for PhD

A more detailed look at Treatment Effects & Matching

Michael E. Kummer

Theoretical Slide Set 7, based on Materials by S. Kastoryano

NovaSBE, OTIM

INTRODUCTION

- In economics (and policy) researchers may be interested in:
 - ▶ How does university education affect future earnings?
 - ▶ How does early life malnutrition affect schooling outcomes?
 - ▶ How does tax cut affect labor participation?
 - ▶ How does having extensive health insurance affect health care use? (moral hazard)
 - ▶ How does a gun law affect the murder rate?
- The researcher is interested in obtaining the causal effect of participating in some treatment on future outcomes.
- Treatment encompasses many definitions, it might refer to an actual intervention, choice variable, individual behavior, or endogenous variable.
- In economic literature, models for analyzing causal effects called models for treatment/policy/program/impact evaluation.

OVERVIEW OF TODAY'S LECTURE

- Define relevant treatment effect parameters.
- Relate potential outcome framework to structural framework.
- Identification through conditional independence assumptions.
- Regression and matching evaluation methods.

POTENTIAL OUTCOMES MODEL

- Potential Outcome Model (aka. Rubin or Neyman-Rubin causal model) finds roots in Neyman (1923) Msc. thesis.
- Further groundwork in statistics by, Rubin (1974), Holland (1986 review).
- In economics, Heckman one of pioneering researchers on policy evaluation (often referring to Roy model).
- Let D_i be an indicator for receiving treatment ($D_i = 1$) or not ($D_i = 0$).
- Each individual has two potential outcomes, $Y_i^{D=1}$ with treatment and $Y_i^{D=0}$ without treatment. We will also denote the realizations of these random variables as $y_i^{D=1}$ and $y_i^{D=0}$.
- The effect for each individual of participating in the treatment equals

$$\Delta_i = Y_i^{D=1} - Y_i^{D=0}$$

- Since only one of the random variables $Y_i^{D=1}$ and $Y_i^{D=0}$ can be observed, Δ_i will always be an unobserved random variable. The unobserved outcome is the *counterfactual outcome*.

ATE, ATET & ATENT

- *Parameter of interest* depends on the context of the study and the (sub-)population of interest. One often considered is **Average Treatment Effect**,

$$ATE = E[Y_i^{D=1} - Y_i^{D=0}] = E[Y_i^{D=1}] - E[Y_i^{D=0}]$$

- If the selection into treatment is not entirely random and some individuals are more likely to enter treatment then it may be preferable to focus on **Average Treatment Effect on the Treated**

$$ATET = E[Y_i^{D=1} - Y_i^{D=0} | D_i = 1] = E[Y_i^{D=1} | D_i = 1] - E[Y_i^{D=0} | D_i = 1]$$

- Notice that ATE can be decomposed into a weighted average effect on the treated and non-treated.

$$\begin{aligned} ATE &= (E[Y_i^{D=1} | D_i = 1] - E[Y_i^{D=0} | D_i = 1]) \cdot \Pr(D_i = 1) \\ &\quad + (E[Y_i^{D=1} | D_i = 0] - E[Y_i^{D=0} | D_i = 0]) \cdot \Pr(D_i = 0) \\ &= ATET \cdot \Pr(D_i = 1) + ATENT \cdot \Pr(D_i = 0) \end{aligned}$$

SOCIAL EXPERIMENTS

- In field experiments, treatment assignment is randomized across individuals.

$$(Y_i^{D=1}, Y_i^{D=0}) \perp D_i$$

- Treatment assignment is statistically independent of potential outcomes, which solves the problem of self-selection.

$$\begin{aligned} ATE &= E[Y_i^{D=1}] - E[Y_i^{D=0}] = E[Y_i^{D=1} | D = 1] - E[Y_i^{D=0} | D = 0] \\ &= E[Y_i | D = 1] - E[Y_i | D = 0] \end{aligned}$$

- This is because treated and non-treated are random sub-samples of population,

$$\begin{aligned} E[Y_i^{D=1}] &= E[Y_i^{D=1} | D_i = 1] = E[Y_i^{D=1} | D_i = 0] \\ E[Y_i^{D=0}] &= E[Y_i^{D=0} | D_i = 1] = E[Y_i^{D=0} | D_i = 0] \end{aligned}$$

- This also implies $ATE = ATET = ATENT$.

QUANTILE TREATMENT EFFECTS

- Average treatment effect (on the treated) only focusses on mean.
- Even if mean treatment effects are zero, it might be that for some individuals treatment effects are positive.
- Knowing effects at different quantiles may allow more efficient targeting of policy.
- Effect at quantile q of distribution of the outcome $F(\cdot)$ given by [Quantile Treatment Effect](#)

$$QTE_{qi} = F_{Y_i^{D=1}}^{-1}(q^{D=1}) - F_{Y_i^{D=0}}^{-1}(q^{D=0})$$

where $q^{D=1}$ is quantile in treated distribution $F_{Y_i^{D=1}}$ at which non-treated individuals at $q^{D=0}$ in $F_{Y_i^{D=0}}$ would have been located had they received treatment (and vice versa for treated).

- But to obtain this effect we must assume treatment does not change an individual's quantile in the distribution...which is a quite strong assumption.

QUANTILE TREATMENT EFFECTS

- If we are not willing to make strong assumption of quantile invariance then we still can define how outcome at a given quantile of the treatment changes due to treatment

$$QTE_q = F_{Y|D=1}^{-1}(q) - F_{Y|D=0}^{-1}(q)$$

- For example, if Y is education and $q = 0.5$ then treatment effect we estimate represents difference in median education (*not* individuals at median under $D = 0$).

SELECTION PROBLEM IN OBSERVATIONAL STUDIES

- Main problem in observational studies: treatment participation is often not independent of potential outcomes, **individuals self-select into treatment**.
- If there is self-selection into the treatment,

$$E[Y_i^{D=1}] \neq E[Y_i^{D=1} | D_i = 1] \quad \text{and} \quad E[Y_i^{D=1}] \neq E[Y_i^{D=1} | D_i = 0]$$

- Example: Unemployed training program
 - ▶ Unemployed workers who are very motivated to find work are more likely to participate in job training programs.
 - ▶ Because these individuals are motivated to find work their unemployment duration (both potential outcomes) will probably be lower than the potential outcomes of less motivated unemployed whether they receive treatment or not. workers.
 - ▶ So if we compare observed unemployment duration outcome for treated who are motivated to unemployment duration outcome of non-treated who are not motivated, our treatment estimate will suffer from *selection bias* which is simply a form of omitted variable bias.

IDENTIFICATION WITHOUT SOCIAL EXPERIMENTS

- Social experiments are less common in economics than in other sciences, such as biology and medicine.
- Without social experiment $E[Y_i^{D=0} | D_i = 1]$ and $E[Y_i^{D=1} | D_i = 0]$ are unobserved.
- Without social experiments, researcher must justify that, perhaps given a set of covariates, there exist comparable individuals besides the fact that some randomly received treatment and others did not.
- To identify *ATE* without social experiment must make additional assumptions.
 - 1 Conditional Independence Assumption
 - 2 Common Support Assumption
 - 3 (Stable Unit Treatment Value Assumption)

IDENTIFICATION: CONDITIONAL INDEPENDENCE ASSUMPTION FOR ATE

- **Conditional Independence Assumption:** after conditioning on \mathbf{x}_i , D_i is as good as randomly assigned,

$$(Y_i^{D=0}, Y_i^{D=1}) \perp D_i | X_i$$

So no self-selection on unobservables.

$$\Pr(D_i = 1 | X_i = \mathbf{x}, Y_i^{D=0} = y_i^{D=0}, Y_i^{D=1} = y_i^{D=1}) = \Pr(D_i = 1 | X_i = \mathbf{x})$$

IDENTIFICATION: COMMON SUPPORT ASSUMPTION FOR ATE

- Identification also requires **Common Support Assumption** (overlap assumption):

$$0 < \Pr(D_i = 1 | X_i = \mathbf{x}) < 1 \quad \text{or simply} \quad 0 < \Pr(D_i = 1 | X_i) < 1$$

- This assumption says there are a sufficiently large number of individuals for all \mathbf{x}_i and there exist both treated and untreated individuals with these characteristics.
- Common Support Assumption can be tested, Conditional Independence Assumption cannot.
- Common support assumption can be tested simply by investigating if for all values of \mathbf{x}_i both treated and untreated individuals exist (see also Black & Smith (Jometrics, 2004) for a graphical test).

IDENTIFICATION: STABLE UNIT TREATMENT VALUE ASSUMPTION

- **Stable Unit Treatment Value Assumption** (SUTVA): treatment participation of one/some units does not affect the potential outcomes of other individuals (or themselves).
- ① Spillover effects: non-treated individuals benefit from treatment (same classroom).
- ② Substitution: non-treated individuals seek alternative treatment (parents find substitute outside classes).
- ③ Hawthorne effect: individuals behave differently in experiment (teachers grade differently during experiment).

ASSUMPTIONS FOR IDENTIFICATION OF ATET

- Non-experimental methods do not rule out selection bias, they simply balance selection bias for the treated and non-treated.
- If we are only interested in ATET then assumptions required for identification are less strong.
- CIA for ATET: Even if some people select into treatment, no one selects out of treatment conditional on X_i .

$$Y_i^{D=0} \perp D_i | X_i$$

- This can be understood as assuming that hypothetically the controls who we match to the treated based on observables are the same as the treated besides that they were, for some reason, unaware about the possibility to enter treatment or the possibility of entering treatment was unavailable.
- We exclude for instance that non-treated status results from an a priori cost-benefit analyses of treatment by individual.
- This is also known as *Selection on observables* or *ignorability of treatment conditional on observables* or *unconfoundedness* and can be rewritten as,

$$F(Y_i^{D=0} | X_i = \mathbf{x}, D_i = 1) = F(Y_i^{D=0} | X_i = \mathbf{x}, D_i = 0)$$

ASSUMPTIONS FOR IDENTIFICATION OF ATET

- CSA now only needs to find matches for treated individuals.

$$Pr(D_i = 1 | X_i) < 1$$

- No one can have characteristics which imply always receiving treatment.

SELECTION TO TREATMENT IN STRUCTURAL MODEL: BALANCING

- Consider structural models,

$$Y_i^{D=0} = E[Y_i^{D=0} | D_i = 0, \mathbf{x}_i] + u_i^{D=0}$$

$$Y_i^{D=1} = E[Y_i^{D=1} | D_i = 1, \mathbf{x}_i] + u_i^{D=1}$$

- When we have selection to treatment, estimation of ATET does not rule out selection bias, so we do not impose

$$E[u_i^{D=0} | D_i = 1, \mathbf{x}_i] = 0$$

- Instead we assume that conditioning on covariates we balance selection bias

$$E[u_i^{D=0} | D_i = 1, \mathbf{x}_i] = E[u_i^{D=0} | D_i = 0, \mathbf{x}_i]$$

- So the zero conditional mean assumption can be violated.
- For ATET we need similar assumptions on $u_i^{D=1}$ and for ATE we need these assumptions on $u_i^{D=1}$ and $u_i^{D=0}$.

METHODS FOR ESTIMATING TREATMENT EFFECTS

Treatment effect literature provides wide range of quite different estimators, many of which are regularly used in empirical work.

- 1 Field (or Social) experiments.
- 2 Regression (including factor models).
- 3 Matching.
- 4 Regression discontinuity.
- 5 Instrumental variable.
- 6 Control Functions.
- 7 Difference-in-difference.
- 8 Nonparametric bounds.
- 9 Timing-of-events.
- 10 Structural estimation (Roy-type models).

DIFFERENCE-IN-MEANS ESTIMATOR

- With unconditional randomization from social experiment, $E[Y_i^{D=1} | D_i = 1]$ and $E[Y_i^{D=0} | D_i = 0]$ can be estimated by their sample means,

$$E[Y_i^{D=1} | D_i = 1] = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} \quad \text{and} \quad E[Y_i^{D=0} | D_i = 0] = \frac{\sum_{i=1}^n (1 - D_i) Y_i}{\sum_{i=1}^n (1 - D_i)}$$

with $D_i = 0, 1$ and $Y_i = (1 - D_i) Y_i^{D=0} + D_i Y_i^{D=1}$ which are always observed.

- The resulting estimator for the treatment effects is called the **difference-in-means** estimator,

$$\widehat{ATE} = \widehat{ATE_T} = \frac{\sum_{i=1}^n D_i Y_i}{\sum_{i=1}^n D_i} - \frac{\sum_{i=1}^n (1 - D_i) Y_i}{\sum_{i=1}^n (1 - D_i)}$$

- The difference-in-means estimator does not impose any structure on the model.

ASYMPTOTIC DISTRIBUTION OF ATE ESTIMATOR

- Assuming unconfoundedness, overlap and some smoothness assumptions on the conditional expectations of potential outcomes Hahn (1998) shows,

$$\sqrt{n}(\widehat{\Delta} - \Delta) \xrightarrow{d} \mathcal{N}(0, V_{\Delta})$$

- where for $\sigma_d^2(\mathbf{x}_i) = \text{Var}(Y_i^{D=d} | X_i = \mathbf{x})$ we can show that,

$$V_{\Delta} \geq E \left[\frac{\sigma_1^2(\mathbf{x}_i)}{p(\mathbf{x}_i)} + \frac{\sigma_0^2(\mathbf{x}_i)}{1-p(\mathbf{x}_i)} + (\widehat{\Delta}(\mathbf{x}_i) - \Delta)^2 \right]$$

- Hahn also shows that asymptotically linear estimators exist that achieve the efficiency bound.

HETEROGENEOUS TREATMENT EFFECTS

- However, ATE may not be only parameter of interest if individuals respond differently to treatment.
- Individuals with different characteristics x_i have different treatment effects
- In such cases, we may want to evaluate treatment effects given a covariate level,

$$ATE(x) = E[Y_i^{D=1} - Y_i^{D=0} | x_i = x] = E[Y_i^{D=1} | x_i = x] - E[Y_i^{D=0} | x_i = x]$$

- In case of a social experiments D_i is also independent of x_i , so
 $ATE(x) = ATET(x) = ATENT(x)$
- ...so why do we often see x_i included in field experiment regressions?
- If x_i discrete and low dimensional can stratify and apply the difference-in-means estimator.
- Difference-in-means becomes inefficient if x_i includes continuous variables or if the stratified samples become small.

LINEAR REGRESSION MODEL

- Alternatively, can specify a linear regression model (here only with one discrete covariate x_i), which can be estimated by OLS

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 D_i x_i + u_i$$

- The linear regression model imposes stronger functional form assumption than difference in means.
- Since distribution of x_i is similar in the treatment and control group we have,

$$ATE(x) = \beta_2 + \beta_3 x \quad \text{and} \quad ATE = \beta_2 + \beta_3 E[x_i]$$

- But what happens if we no longer have a social experiment and the set of conditioning variables in the CSA can not be saturated in the model? Do the parameters still represent the ATE ?

REGRESSION ESTIMATION OF TREATMENT EFFECTS

- Consider first a simple linear regression model,

$$Y_i = X_i' \beta + \delta D_i + u_i$$

- δ captures treatment effect although it is not always clear in practice whether this is *ATE* or *ATET* or something else.
- This regression is most often used when randomization of D_i is unconditional since in that case we know $ATE = ATET$.
- Sometimes also used within the context of factor models.
- This simple specification is unlikely to adequately capture correlations between covariates, treatment and unobservables.
- Errors are therefore unlikely to be balanced.

TWO STEP FITTED REGRESSION

- Regression approach with weaker specification assumption is to:

- 1 Estimate model for $E[Y_i^{D=1} | D_i = 1, X_i = \mathbf{x}]$ by linear regression $E[Y_i | D_i = 1, \mathbf{x}_i]$. Using estimated coefficients generate predicted (fitted) values $\hat{Y}_i^{D=1}$ for all treated and non-treated individuals.
- 2 Estimate model for $E[Y_i^{D=0} | D_i = 0, X_i = \mathbf{x}]$ by linear regression $E[Y_i | D_i = 0, \mathbf{x}_i]$. Using estimated coefficients generate predicted (fitted) values $\hat{Y}_i^{D=0}$ for all treated and non-treated individuals.
- 3 Compute

$$ATE = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i^{D=1} - \hat{Y}_i^{D=0}$$

$$ATE_T = \left(\frac{1}{n} \sum_{i=1}^n D_i (\hat{Y}_i^{D=1} - \hat{Y}_i^{D=0}) \right) / \frac{1}{n} \sum_{i=1}^n D_i$$

- 4 Compute standard errors by bootstrapping this procedure.

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Regression estimators can be sensitive to differences in the covariate distributions for treated and control units.
- If distribution of covariates for treated different than that for controls then fitted values can be sensitive to changes in specification.
- Matching also fits counterfactual outcomes but in a way less sensitive to specification.
- Idea is to find for each treated unit (a set of) 'similar' non-treated individuals (and vice-versa for non-treated).
- Assuming conditional independence holds, we can then estimate treatment effect by comparing outcomes for individuals with similar covariates.

MATCHING ESTIMATORS TO ACCOUNT FOR SELECTION ON OBSERVABLES

- Suppose: n_1 individuals observed to receive treatment and n_0 individuals without treatment

$$ATE_T = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(Y_i \cdot D_i - \sum_{j=1}^{n_0} W(i,j) \cdot Y_j \cdot (1 - D_j) \right)$$

- $W(i,j)$, where $\sum_{j=1}^{n_0} W(i,j) = 1$ for all i , weights non-treated individuals in such a way to construct the counterfactual for individual i in the treatment group.
- Many types of weights to specify $W(i,j)$. Lead to different matching estimators: Nearest Neighbour Matching, Kernel Matching, one-to-one (with or without replacement), etc.
- If $W(i,j) = 1/n_0$, then we have the difference-in-means estimator.

MATCHING BASED ON PROPENSITY SCORE

- Finding identical individuals in the treatment and control group suffers from the curse of dimensionality.
- Exact matching on covariates is often not feasible.
- One commonly used approach to reduce dimensionality problem is to match based on propensity score.
- Propensity score** is probability of entering treatment conditional on covariates:

$$p(\mathbf{x}_i) = \Pr(D_i = 1 | \mathbf{x}_i = \mathbf{x})$$

- Rosenbaum and Rubin (Biometrika, 1983) show that CIA for ATE and ATET imply

$$(Y_i^{D=1}, Y_i^{D=0}) \perp D_i | p(\mathbf{x}_i) \quad \text{and} \quad Y_i^{D=0} \perp D_i | p(\mathbf{x}_i)$$

- Instead of matching on all \mathbf{x}_i , we can match on $p(\mathbf{x}_i)$.

MATCHING BASED ON PROPENSITY SCORE: ESTIMATION

- ① Estimate binary model $Pr(D_i = 1 | X_i = \mathbf{x})$ nonparametrically or for example with Logit model.
- ② Next compute the $\hat{p}(\mathbf{x}_i)$ for all i .
- ③ Use nearest-neighbor matching, kernel-matching, etc. on $\hat{p}(\mathbf{x}_i)$ to match treated to an (weighted set of) individual(s).
- ④ In large samples, different estimators tend to be very similar.
- ⑤ After using propensity score matching, researchers often compare the distribution of X_i -variables in the treatment and constructed control group.
- ⑥ Check if X_i -variables are balanced, if not, trim for the group of treated which have overlapping non-treated (this will again affect the subpopulation you define as 'treated').

PROPENSITY SCORE WEIGHTING

- Another approach instead of matching based on the propensity score is weighting on the propensity score (e.g. Hirano, Imbens and Ridder, Ecetra 2003).
- If the true propensity score $p(\mathbf{x}_i)$ were known the ATE and ATET would be given by,

$$ATE = \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{p(\mathbf{x}_i)} - \frac{(1 - D_i) Y_i}{(1 - p(\mathbf{x}_i))}$$

$$ATET = \frac{\frac{1}{n} \sum_{i=1}^n p(\mathbf{x}_i) \left(\frac{D_i Y_i}{p(\mathbf{x}_i)} - \frac{(1 - D_i) Y_i}{(1 - p(\mathbf{x}_i))} \right)}{\frac{1}{n} \sum_{i=1}^n p(\mathbf{x}_i)}$$

PROPNESITY SCORE WEIGHTING

$$ATE = \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{p(\mathbf{x}_i)} - \frac{(1 - D_i) Y_i}{(1 - p(\mathbf{x}_i))}$$

- Since $p(\mathbf{x}_i)$ is unknown, estimating the above by replacing $p(\mathbf{x}_i)$ by some estimator $\hat{p}(\mathbf{x}_i)$ is not necessarily efficient.
- Why? Consider the treated. In the population it is clear that $E[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{D_i}{p(\mathbf{x}_i)}] = E[E[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{D_i}{p(\mathbf{x}_i)} | X_i = \mathbf{x}]] = 1$ holds.
- But in the sample $\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{D_i}{p(\mathbf{x}_i)} = 1$ will not exactly hold because $\text{Var}[\frac{1}{n_1} \sum_{i=1}^{n_1} \frac{D_i}{p(\mathbf{x}_i)}] > 0$ (and similar arguments for the untreated).
- Essentially, by using the true propensity score we are attributing too much weight to some observations while attributing too little to others within the sample.
- Solution is simply to use 'incorrect' finite sample propensity score estimates $\tilde{p}(\mathbf{x}_i)$ which converge to $p(\mathbf{x}_i)$ as $n \rightarrow \infty$.
- Hirano, Imbens and Ridder (2003) propose a series logit estimator where order of series terms is a function of the sample size..

INTUITION BEHIND PROPENSITY SCORE WEIGHTING

- A relevant question you may ask is why this probability of treatment ends up in the denominator?
- Recall that by construction $\frac{1}{n_1} \sum_{i=1}^{n_1} p(\mathbf{x}_i) \geq \frac{1}{n_0} \sum_{j=1}^{n_0} p(\mathbf{x}_j)$ which simply says that those individuals observed to receive treatment have a higher propensity to treatment based on observable variables.
- Now suppose we observe a treated subject with propensity score of 0.2.
- So we observe that treatment occurred despite this individual having relatively low propensity to treatment based on covariates.
- This individual therefore carries a relatively large amount of information concerning the effect of the treatment for the non-treated.
- More generally, treated observations with low propensity score and un-treated observations with high propensity score carry relatively more information about the ATE.
- Using $\tilde{p}(\mathbf{x}_i)$ rather than $p(\mathbf{x}_i)$ in this sense produces the appropriate representation of each observation in the hypothetical post-interventional data given the observed sample.

TREATMENT EFFECT METHODS SO FAR

Let's recap two regression methods we have used so far to estimate ATE and ATET:

1 Regression 1: linear regression model

$$Y_i = X_i' \beta + \delta D_i + u_i$$

δ captures a treatment effect (in case of randomized experiment, $ATE = ATET$).

2 Regression 2: Two step fitted regression

$$ATE = \frac{1}{n} \sum_{i=1}^n \hat{Y}_i^{D=1} - \hat{Y}_i^{D=0}$$
$$ATET = \left(\frac{1}{n} \sum_{i=1}^n D_i (\hat{Y}_i^{D=1} - \hat{Y}_i^{D=0}) \right) / \frac{1}{n} \sum_{i=1}^n D_i$$

TREATMENT EFFECT METHODS SO FAR

Let's recap two matching methods we have used so far to estimate ATE and ATET:

❶ (Propensity Score) Matching:

- ▶ Calculate *ATE* and *ATET* by creating a synthetic un-treated observation for each treated observation and a synthetic treated observation for each un-treated observation.
- ▶ Then for $i = 1, \dots, n_1$ treated and $j = 1, \dots, n_0$ controls we have

$$ATET = \frac{1}{n_1} \sum_{i=1}^{n_1} \left(Y_i \cdot D_i - \sum_{j=1}^{n_0} W(i, j) Y_j \cdot (1 - D_j) \right)$$

$$ATE = \frac{1}{n_0 + n_1} \sum_{i=1}^{n_1} \left(Y_i \cdot D_i - \sum_{j=1}^{n_0} W(i, j) Y_j \cdot (1 - D_j) \right) +$$

$$\frac{1}{n_0 + n_1} \sum_{j=1}^{n_0} \left(\sum_{i=1}^{n_1} W(j, i) Y_i \cdot D_i - Y_j \cdot (1 - D_j) \right)$$

❷ Propensity Score Weighting:

TREATMENT EFFECT METHODS SO FAR

Let's recap two matching methods we have used so far to estimate ATE and ATET:

- ① Propensity Score Matching:
- ② Propensity Score Weighting:

► For some normalized estimate $\tilde{p}(\mathbf{x}_i)$ of $p(\mathbf{x}_i)$ we can obtain ATE and ATET by:

$$ATE = \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\tilde{p}(\mathbf{x}_i)} - \frac{(1 - D_i) Y_i}{(1 - \tilde{p}(\mathbf{x}_i))}$$

$$ATET = \frac{\frac{1}{n} \sum_{i=1}^n \tilde{p}(\mathbf{x}_i) \left(\frac{D_i Y_i}{\tilde{p}(\mathbf{x}_i)} - \frac{(1 - D_i) Y_i}{(1 - \tilde{p}(\mathbf{x}_i))} \right)}{\frac{1}{n} \sum_{i=1}^n \tilde{p}(\mathbf{x}_i)}$$

HOW CAN WEIGHTING AND OTHER METHODS BE COMPARED?

- A relevant question that should come out of the previous overview is why weighting vs. non-weighting methods would both estimate same treatment effects?
- To answer this let's consider simple situation where we have $Y = 0, 1$ binary, $D = 0, 1$ binary and some characteristic X (from J. Pearl blog).
- Let's say X is discrete and we have the *pre-treatment* joint distribution $\Pr(Y = 1, D = d, X = x)$.
- This can be decomposed into a product of three conditional probabilities. In the case of $D = 1$ (treatment actually occurs in the future):

$$\Pr(Y = 1, D = 1, X = x) = \Pr(Y = 1 | D = 1, X = x) \cdot \Pr(D = 1 | X = x) \cdot \Pr(X = x)$$

- ① $\Pr(Y = 1 | D = 1, X = x)$ describes how the outcome depends on treatment and covariate X .
- ② $\Pr(D = 1 | X = x)$ describes how subjects choose treatment prior to treatment.
- ③ $\Pr(X = x)$ describes the prior distribution of the covariates.

HOW CAN WEIGHTING AND OTHER METHODS BE COMPARED?

- However, to calculate the treatment effects we are interested in comparing the *post-treatment* distributions $E[Y^{D=d}] = \sum_x \Pr^*(Y = 1, D = d, X = x)$ for $D = 0, 1$.
- Let's decompose again $\Pr^*(Y = 1, D = d, X = x)$ when $D = 1$ (treatment has actually occurred):

$$\Pr^*(Y = 1, D = 1, X = x) = \Pr^*(Y = 1 | D = 1, X = x) \cdot \Pr^*(D = 1 | X = x) \cdot \Pr^*(X = x)$$

- Which of the three mechanisms remains invariant?
- Clearly, $\Pr^*(X = x) = \Pr(X = x)$ as long as there is no attrition and X does not contain variables endogenous to treatment.
- Also $\Pr^*(Y = 1 | D = 1, X = x) = \Pr(Y = 1 | D = 1, X = x)$ because X is assumed to be complete set of confounders, so, Y depends on D and X plus random noise (error) that is not affected by the treatment.

HOW CAN WEIGHTING AND OTHER METHODS BE COMPARED?

- However, the treatment mechanism changes drastically since now $\Pr^*(D = 1|X = x) = 1$ because everyone gets the treatment.
- As a result, our post-treatment distribution of interest is equal to the product:

$$\Pr^*(Y = 1, D = 1, X = x) = \Pr(Y = 1|D = 1, X = x) \cdot \Pr(X = x)$$

- Or the function

$$\Pr^*(Y = 1, D = 1, X = x) = \frac{\Pr(Y = 1, D = 1, X = x)}{\Pr(D = 1|X = x)}$$

- First perspective leads to estimation by regression, stratification, or (propensity score) matching, while the second leads to Inverse Probability Weighting.
- The asymptotic equivalence of the two approaches is assured by the equality of the two equations for $\Pr^*(Y = 1, D = 1, X = x)$.

ROBUSTNESS OF SPECIFICATION VS IDENTIFICATION

- Some say that the choice of methods should be guided by which process you as a researcher think you can model best:
 - ① If you know assignment mechanism on covariates and this assignment mechanism is unknown to agents, then propensity weighting may be reasonable.
 - ② If you know the outcome mechanism given the treatment or think you can adjust selection bias adequately in the outcome equation, then a regression type approach may be reasonable.
- But if you don't have a strong knowledge of the mechanism inducing randomization then you have an identification problem.
- Without clear identification, none of the estimators can guarantee you are estimating a meaningful parameter.
- If you know the covariates determining selection but are unsure (or they are high dimensional/continuous) and you fear model misspecification then can use doubly robust estimator.
- In the assignment you will explore these differences further both in theory and practice.

DOUBLY ROBUST ESTIMATION

- Doubly robust estimator combines regression and weighting and is consistent even if one of outcome or treatment mechanisms is misspecified.
- Consider the weighted model,

$$\frac{Y_i}{w_i} = \frac{1}{w_i}(x_i'\beta + \delta D_i + u_i) = \frac{1}{w_i}(m(\mathbf{x}) + u_i)$$

- We can combine inverse probability weighting and regression by using weights

$$1/w_i = \sqrt{\frac{D_i}{p(x_i)} + \frac{1-D_i}{1-p(x_i)}}$$

- Least squares estimation results in the doubly robust estimator,

$$\begin{aligned} ATE &= \hat{Y}^{D=1} - \hat{Y}^{D=0} \\ &= \frac{1}{n} \sum_{i=1}^n \frac{D_i Y_i}{\tilde{p}(\mathbf{x}_i)} - \frac{D_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)} \hat{m}_1(\mathbf{x}) \\ &\quad - \frac{1}{n} \sum_{i=1}^n \frac{(1-D_i) Y_i}{(1-\tilde{p}(\mathbf{x}_i))} + \frac{D_i - \tilde{p}(\mathbf{x}_i)}{1-\tilde{p}(\mathbf{x}_i)} \hat{m}_0(\mathbf{x}) \end{aligned}$$

DOUBLY ROBUST ESTIMATOR PROPERTY

- Appealing property of doubly robust estimator is that it remains consistent as long as one of the propensity score or the outcome regression are correctly specified.
- To see this, consider the term $\widehat{Y}_i^{D=1}$ and rewrite

$$\begin{aligned} E[\widehat{Y}_i^{D=1}] &= E\left[\frac{D_i Y_i}{\tilde{p}(\mathbf{x}_i)} - \frac{D_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)} \hat{m}_1(\mathbf{x})\right] \\ &= E\left[Y_i^{D=1} + \frac{D_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)} (Y_i^{D=1} - \hat{m}_1(\mathbf{x}))\right] \\ &= E[Y_i^{D=1}] + E\left[\frac{D_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)} (Y_i^{D=1} - \hat{m}_1(\mathbf{x}))\right] \end{aligned}$$

- We can see that the second term is equal to 0 if

$$E\left[E\left[\frac{D_i - \tilde{p}(\mathbf{x}_i)}{\tilde{p}(\mathbf{x}_i)}\right] \middle| Y_i^{D=1}, \mathbf{x}_i\right] = 0 \quad \text{or} \quad E\left[E\left[Y_i^{D=1} - \hat{m}_1(\mathbf{x})\right] \middle| Y_i^{D=1}, \mathbf{x}_i\right] = 0$$

- The double robust estimator was developed by Robins & Rotnitzky (1995) in the context of missing data. In that context, the inverse probability weighting adjusts for the propensity to be missing in the sample.