

# Topic 3: Selection on Observables

*ECON 5783 – University of Arkansas*

Prof. Kyle Butts

Fall 2024

# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Selection into Treatment

In our last lectures, we have covered the problem with difference-in-means comparisons:

- If units with  $D_i = 1$  have different characteristics on average than units with  $D_i = 0$ , then we are making faulty comparisons

## Selection into Treatment

In our last lectures, we have covered the problem with difference-in-means comparisons:

- If units with  $D_i = 1$  have different characteristics on average than units with  $D_i = 0$ , then we are making faulty comparisons

In the absence of treatment, the control units and the treated units would have different average  $Y_i(0)$  due to differences in characteristics

## Selection into Treatment and Omitted Variable Bias

We have shown that if there is a single problematic covariate,  $X_2$ , that differs between the treated group and control group, our difference-in-means estimator is biased:

$$\begin{aligned}\hat{\tau}_{\text{DIM}} &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \\ &= \tau_{\text{ATT}} + \frac{\beta_2}{1 - \mathbb{P}(D = 1)} (\mathbb{E}[X_2 \mid D = 1] - \mathbb{E}[X_2])\end{aligned}$$

## Selection into Treatment and Omitted Variable Bias

We have shown that if there is a single problematic covariate,  $X_2$ , that differs between the treated group and control group, our difference-in-means estimator is biased:

$$\begin{aligned}\hat{\tau}_{\text{DIM}} &= \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] \\ &= \tau_{\text{ATT}} + \frac{\beta_2}{1 - \mathbb{P}(D = 1)} (\mathbb{E}[X_2 | D = 1] - \mathbb{E}[X_2])\end{aligned}$$

- The ATT estimate reflects the impact that differences in  $X_2$  have on the outcome  $\beta_2$

## Controls

This raises the simple question: what if we compare treated and control units with the same values of  $X_2$ ?

- Subset the control units to those that “look like” the treated units

## Controls

This raises the simple question: what if we compare treated and control units with the same values of  $X_2$ ?

- Subset the control units to those that “look like” the treated units

This is a reasonable strategy in some settings and understanding it is the goal of this topic

# Conditional Independence Assumption

The **Conditional Independence Assumption** is given by

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- This says that for units with the same value of  $\mathbf{X}_i = x$ , treatment is independent of the potential outcomes (i.e. “randomly assigned”)

# Conditional Independence Assumption

The **Conditional Independence Assumption** is given by

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

- This says that for units with the same value of  $\mathbf{X}_i = x$ , treatment is independent of the potential outcomes (i.e. “randomly assigned”)

This is like a randomized control trial except you can only compare treated and control units with the same  $\mathbf{X}_i$

## Example 1

### *Effect of attending a prestigious school*

Dale and Krueger (2002, QJE) estimate the effect of attending a more selective college on future earnings

- Whether or not you get accepted into a more selective college is not randomly assigned

## Example 1

### *Effect of attending a prestigious school*

Dale and Krueger (2002, QJE) estimate the effect of attending a more selective college on future earnings

- Whether or not you get accepted into a more selective college is not randomly assigned

Using data on the full list of schools that students were accepted to, they argue that for students who got into the same schools, whether they went to the more selective college is random (independent of potential outcomes)

## Example 2

### *Winning the Lottery*

Imbens, Rubin, and Sacerdote (2001, AER) estimate the impact of unearned income on labor supply using the lottery

- While winning the lottery is random *among those that play*, playing the lottery is not randomly assigned.
- Even among players, some play more frequently
- Comparing lottery winners labor supply to the population would be biased since *lottery players look different*

## Example 2

### *Winning the Lottery*

Imbens, Rubin, and Sacerdote (2001, AER) estimate the impact of unearned income on labor supply using the lottery

- While winning the lottery is random *among those that play*, playing the lottery is not randomly assigned.
- Even among players, some play more frequently
- Comparing lottery winners labor supply to the population would be biased since *lottery players look different*

They argue that conditional on playing and winning in the lottery ( $X_i$ ), winning a small amount (control) or a large amount (treatment) is randomly assigned

## Example 3

### *Charter Schools*

Abdulkadiroğlu, Angrist, and Pathak (2014, ECTA) estimate the impact of attending an elite “exam school” in NYC on students’ success.

- These schools accept students with very high scores on a centralized exam

## Example 3

### *Charter Schools*

Abdulkadiroğlu, Angrist, and Pathak (2014, ECTA) estimate the impact of attending an elite “exam school” in NYC on students’ success.

- These schools accept students with very high scores on a centralized exam

However, these schools are over-subscribed to and have to create admission cutoffs

- Compare students just on either end of the cutoff score. Since the test scores are likely noisy, these students are randomly assigned to the best schools or slightly lower quality schools

## Example 4

### *College Athletic Success*

Anderson (2017, RESTAT) estimates the effect of having a successful college athletics program help colleges (say recruit better students)

- Whether or not a school is good at athletics is not randomly assigned

Uses gambling bookmakers' "spreads" as a measure of predicted odds of a school winning (in football) games. They compare schools with the same odds of winning; those that won more than expected to those that won fewer

## Example 5

### *Macroeconomic Policy Shocks*

Angrist, Jordà, and Kuersteiner (2016, JASA) estimate the impact of unexpected changes to interest rates on the economy

- The Fed makes decisions on interest rates based on economic conditions (reverse causality)

The prices of contracts in the futures markets reflect the expected change in interest rates, so they compare interest rate changes with the same expected change but different unexpected changes

## Block-randomized Experiments

Say you run an experiment where you assign treatment to men randomly with probability 25% and to women randomly with probability 75%.

- By randomization, we have  $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid F_i$  where  $F_i$  is an indicator for being a female

## Block-randomized Experiments

Say you run an experiment where you assign treatment to men randomly with probability 25% and to women randomly with probability 75%.

- By randomization, we have  $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid F_i$  where  $F_i$  is an indicator for being a female

Say  $\mathbb{E}[Y_i(0) \mid F_i = 1] \neq \mathbb{E}[Y_i(0) \mid F_i = 0]$ .

- Then since treatment is correlated with gender,  $Y_i(0) \not\perp\!\!\!\perp D_i$ . Even though we randomize for each gender, we do not have (unconditional) independence

## Clustered Experiments

Our difference-in-means estimator using the whole population will be biased.

- However, doing difference-in-means separately for male and females is valid (conditional independence holds!)

What does each difference-in-means estimator estimate?

# Clustered Experiments

What does each difference-in-means estimator estimate? Under conditional independence

$$\tau_{\text{DIM, Female}} = \mathbb{E}[Y_i \mid D_i = 1, F_i = 1] - \mathbb{E}[Y_i \mid D_i = 0, F_i = 1]$$

# Clustered Experiments

What does each difference-in-means estimator estimate? Under conditional independence

$$\begin{aligned}\tau_{\text{DIM, Female}} &= \mathbb{E}[Y_i \mid D_i = 1, F_i = 1] - \mathbb{E}[Y_i \mid D_i = 0, F_i = 1] \\ &= \mathbb{E}[Y_i(1) - Y_i(0) \mid F_i = 1] = \tau(F_i = 1)\end{aligned}$$

We estimate the conditional average treatment effect by gender!

# Clustered Experiments

How do we estimate the overall average treatment effect?

$$\tau_{ATE} = \mathbb{E}[\tau_i] = \mathbb{E}[\tau(F_i)]$$

$$= \tau(F_i = 1) * \mathbb{P}(F_i = 1) + \tau(F_i = 0) \mathbb{P}(F_i = 0)$$

- Weighted average of the CATE with weights proportional to group size

## Generalization

In more general settings where  $X_i$  is (possibly) continuous and multi-dimensional, we can still do the same procedure:

- For each value of  $x$ , calculate difference in means of treated and control units with  $X_i = x$
- Integrate (average) the difference-in-means estimates using the distribution of  $X$

## Generalization

In more general settings where  $\mathbf{X}_i$  is (possibly) continuous and multi-dimensional, we can still do the same procedure:

- For each value of  $x$ , calculate difference in means of treated and control units with  $\mathbf{X}_i = \mathbf{x}$
- Integrate (average) the difference-in-means estimates using the distribution of  $X$

$$\begin{aligned}\tau_{ATE} &= \int_{\mathbf{x} \in \mathbb{X}} \tau(\mathbf{X}_i = \mathbf{x}) d\mathbb{P}(\mathbf{X} = \mathbf{x}) \\ &= \int_{\mathbf{x} \in \mathbb{X}} \mathbb{E}[Y_i | D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i | D_i = 0, \mathbf{X}_i = \mathbf{x}] d\mathbb{P}(\mathbf{X} = \mathbf{x})\end{aligned}$$

## Generalization

If we want the average treatment effect on the treated, we average over the distribution of  $\mathbf{X}$  for the treated group

$$\begin{aligned}\tau_{\text{ATT}} &= \int_{\mathbf{x} \in \mathbb{X}} \tau(\mathbf{X}_i = \mathbf{x}) d\mathbb{P}(\mathbf{X} = \mathbf{x} \mid D_i = 1) \\ &= \int_{\mathbf{x} \in \mathbb{X}} \mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i = \mathbf{x}] d\mathbb{P}(\mathbf{X} = \mathbf{x} \mid D_i = 1)\end{aligned}$$

## Example 2

### *Attending College*

Say we want to know the effect of college on future earnings in an observational study.

We might not believe that attending college  $D_i$  is randomly assigned to workers

- But we might think that for workers with the same high-school GPA, same family income, and same parental education status ( $X_i$ ), whether or not a worker attends college is independent of outcomes

## What drives treatment after conditioning on $X_i$

To be clear, *something* must be causing some units to sign up for treatment and others not to

- Maybe some teacher encouraged person A to apply but person B was not
  - Since this is probably unrelated to future earnings, this is in support of conditional independence assumption

## What drives treatment after conditioning on $X_i$

To be clear, *something* must be causing some units to sign up for treatment and others not to

- Maybe some teacher encouraged person A to apply but person B was not
  - Since this is probably unrelated to future earnings, this is in support of conditional independence assumption
- Or maybe how many college-going classmates you impacts whether or not you go
  - Since these peer effects probably impact future earnings, this is not in support of conditional independence assumption

## Using this method

There is always going to be debate with this method on what variables are missing in  $\mathbf{X}_i$



Apollo

X includes a lot of important factors that drive selection into treatment!



von Karma

I think there are other omitted variables that drive selection into treatment!!

## Using this method

There is always going to be debate with this method on what variables are missing in  $X_i$

- Author: " $X_i$  includes a lot of important factors that drive selection into treatment!"
- Reviewer: "I think there are other omitted variables that drive selection into treatment!!"

I think this is one reason why this method is not more popular; your luck in publishing depends on whether or not your referee "believes" you have included the right variables in  $X_i$

## Military Service and Future Earnings

Angrist (1998) studies how military service affects future earnings

- Non-random selection into volunteering for the military
- Matches on  $X$  = race, application year, years of schooling, Armed Forces Qualification Test score group, and year of birth

Compares each veteran to a non-veteran worker with the same exact characteristics

# Military Service and Future Earnings

Angrist (1998) studies how military service affects future earnings

- Non-random selection into volunteering for the military
- Matches on  $X$  = race, application year, years of schooling, Armed Forces Qualification Test score group, and year of birth

Compares each veteran to a non-veteran worker with the same exact characteristics

- Might not be perfect, but does a better job at alleviating concerns about selection into treatment
- I think the AFQT score helps quite a bit

# Why selection on observables is hard

## *Evictions*

Collinson, Humphries, Mader, Reed, Tannenbaum, and Van Dijk (2023, QJE) study the effect of being evicted on future earnings

Eviction courts randomly assign cases to judges of varying 'leniency'

- Some are assigned to harsh judges while others are assigned to more lenient ones

# Why selection on observables is hard

## *Evictions*

They first perform an analysis like previous work in the literature:

- Compare evicted and non-evicted individuals of similar demographics: e.g. income, race, highest level of education, gender, marital status, children, age, past criminal record, past job loss, past relationship dissolution, and housing assistance receipt

They find large and significant effects on earnings

# Why selection on observables is hard

## *Evictions*

Second, they show using administrative data on credit-scores, that even conditional on those demographic characteristics, those that get evicted have sharp declines in earnings just prior to eviction

- The evicted people likely would have worse  $Y(0)$  (in the absence of eviction) than those that were not evicted

# Why selection on observables is hard

## *Evictions*

Second, they show using administrative data on credit-scores, that even conditional on those demographic characteristics, those that get evicted have sharp declines in earnings just prior to eviction

- The evicted people likely would have worse  $Y(0)$  (in the absence of eviction) than those that were not evicted

Results are still negative, but significantly smaller

- Even though the observables that previous work has matched on were quite 'rich', there was still a large role for unobservables

# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Reminder of Setup

We observe a sample of  $n$  observations,  $Y$  is outcome,  $\mathbf{X}$  is a vector of control variables,  $D$  is our treatment variable of interest. We assume the conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

## Reminder of Setup

We observe a sample of  $n$  observations,  $Y$  is outcome,  $\mathbf{X}$  is a vector of control variables,  $D$  is our treatment variable of interest. We assume the conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

Our average-treatment effect estimate can be written as the average of CATE (averaged over distribution of  $X$ )

$$\tau_{ATE} = \mathbb{E}[ \mathbb{E}[Y_i(1) \mid \mathbf{X}_i] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i] ]$$

# Conditional Expectation Function

$$\tau_{ATE} = \mathbb{E}[ \mathbb{E}[Y_i(1) | \mathbf{X}_i] - \mathbb{E}[Y_i(0) | \mathbf{X}_i] ]$$

There are two conditional expectation functions in the above terms. Define

$$\mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}] \text{ and } \mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) | \mathbf{X}_i = \mathbf{x}]$$

# Conditional Expectation Function

$$\tau_{ATE} = \mathbb{E}[ \mathbb{E}[Y_i(1) | \mathbf{X}_i] - \mathbb{E}[Y_i(0) | \mathbf{X}_i] ]$$

There are two conditional expectation functions in the above terms. Define

$$\mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) | \mathbf{X}_i = \mathbf{x}] \text{ and } \mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) | \mathbf{X}_i = \mathbf{x}]$$

Then, the above becomes  $\tau_{ATE} = \mathbb{E}[\mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)]$

- If we knew the CEF, then we can estimate this with  $\frac{1}{n} \sum_{i=1}^n \mu_1(\mathbf{X}_i) - \mu_0(\mathbf{X}_i)$

## Estimation of the conditional expectation functions

$$\mu_0(\mathbf{x}) = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] \text{ and } \mu_1(\mathbf{x}) = \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}]$$

Since we have conditional independence assumption:

$$\mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}, D_i = 0] = \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] = \mu_0(\mathbf{x})$$

- We can estimate  $\mu_0(\cdot)$  and  $\mu_1(\cdot)$  using the untreated and treated group, respectively
- We have talked about many different ways to estimate the CEF (e.g. linear regression)

## Linear model for CEF

Say we estimate  $\mu_0(x)$  and  $\mu_1(x)$  using linear regressions of  $Y$  on  $\mathbf{X}$  for the untreated and treated groups respectively:

$$Y_i = \alpha_d + \mathbf{X}'_i \beta_d + u_{id}$$

- The control-sample regression gives us  $\hat{\alpha}_0, \hat{\beta}_0$  and the treated-sample regression gives us  $\hat{\alpha}_1, \hat{\beta}_1$

# Regression Adjustment

## ATE Estimator

The **regression adjustment treatment effect estimators** are given by:

$$\hat{\tau}_{ATE} = \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{\alpha}_1 + \mathbf{X}'_i \hat{\beta}_1)}_{\hat{\mathbb{E}}[Y_i(1)]} - \underbrace{\frac{1}{n} \sum_{i=1}^n (\hat{\alpha}_0 + \mathbf{X}'_i \hat{\beta}_0)}_{\hat{\mathbb{E}}[Y_i(0)]}$$

- Predicting everyone's  $Y_i(1)$  and  $Y_i(0)$  given their  $\mathbf{X}_i$  and taking the difference
- Quality of estimator therefore depends on quality of conditional expectation model

# Regression Adjustment

*ATT Estimator*

$$\hat{\tau}_{\text{ATT}} = \underbrace{\frac{1}{n_1} \sum_{i=1}^n D_i (\hat{\alpha}_1 + \mathbf{X}'_i \hat{\beta}_1)}_{\hat{\mathbb{E}}[Y_i(1) \mid D_i=1]} - \underbrace{\frac{1}{n_1} \sum_{i=1}^n D_i (\hat{\alpha}_0 + \mathbf{X}'_i \hat{\beta}_0)}_{\hat{\mathbb{E}}[Y_i(0) \mid D_i=1]}$$

# Regression Adjustment

For the ATT, remember that the fitted value of a linear regression averaged over the estimation sample is the average of  $Y$  for that sample.

So, we can rewrite the

$$\begin{aligned}\hat{\tau}_{\text{ATT}} &= \frac{1}{n_1} \sum_{i=1}^n D_i (\hat{\alpha}_1 + \mathbf{X}'_i \hat{\beta}_1) - \frac{1}{n_1} \sum_{i=1}^n D_i (\hat{\alpha}_0 + \mathbf{X}'_i \hat{\beta}_0) \\ &= \frac{1}{n_1} \sum_{i=1}^n D_i Y_i - \frac{1}{n_1} \sum_{i=1}^n D_i (\hat{\alpha}_0 + \mathbf{X}'_i \hat{\beta}_0)\end{aligned}$$

- Only relying on the model for  $\mu_0(x)$  to estimate the ATT

# Regression Adjustment via Regression

You can force regression to estimate this for you in a single regression; this makes producing valid standard errors easier

$$Y_i = \alpha_0 + \alpha_1 D_i + \beta_0 \mathbf{X}_i + \beta_1 D_i * (\mathbf{X}_i - \bar{\mathbf{X}}) + u_i$$

- Then our  $\hat{\tau}_{ATE}$  estimate is given by  $\hat{\alpha}_1$
- If you subtract off the average of  $\mathbf{X}$  for the treated group,  $\hat{\alpha}_1$  will be  $\hat{\tau}_{ATT}$

## Coding implementation

In R, use the `fixest` package to estimate the regression with HC1 standard errors

In Stata, use the `teffects ra` command

## Function form of CEF Models

The main difficulty with this estimator is that we need to be confident in our model of

$$\mu_d(\mathbf{x}) = \mathbb{E}[Y_i(d) \mid \mathbf{X}_i = \mathbf{x}]$$

- Can do better if we include polynomials of or bin continuous variables
- Set of indicators for discrete variables
- Consider important interactions between variables

See Imbens and Wooldridge (2009, JEL) review article for more details

# Regression Adjustment in RCTs

When treatment is completely randomly assigned, covariates are balanced *on average*

- For a single draw of the experiment, some covariates may remain unbalanced

# Regression Adjustment in RCTs

When treatment is completely randomly assigned, covariates are balanced *on average*

- For a single draw of the experiment, some covariates may remain unbalanced

It is often common practice to show the difference-in-means estimator and a second regression-adjustment estimate

- Lin (2013) and Nagi and Wooldridge (2021) show that the latter improves efficiency even in an RCT

# Roadmap

Conditional Independence Assumption

Regression Adjustment

**Matching estimators**

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Curse of dimensionality strikes again!

Given the conditional independence assumption (CIA),

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

we can theoretically perform a bunch of different difference-in-means estimators and then aggregate covariate-specific estimates

- This works well when  $X_i$  is a single binary variable

## Curse of dimensionality strikes again!

Given the conditional independence assumption (CIA),

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i$$

we can theoretically perform a bunch of different difference-in-means estimators and then aggregate covariate-specific estimates

- This works well when  $X_i$  is a single binary variable

But as we add more and more covariates, the fewer observations we have for  $\mathbf{X}_i = \mathbf{x}$

- What do we do if there are no treated units or no control units for a given  $\mathbf{x}$ ?

## Finding comparable control units

What do we do if there are no treated units or no control units for a given  $x$ ?

One strategy we could take is to find a control unit for each treated unit that has an  $X_i$  that is “close” to a given treated unit’s  $x$

- For each treated unit  $i$ , find a matching control unit (or units),  $j$ , that has  $X_i \approx X_j$

## Matching estimator

Let  $j(i)$  be the unit that is matched to treated unit  $i$ . Let  $N_1 = \sum_i D_i$  be the number of treated units.

Our treatment effect estimator is given by:

$$\hat{\tau}_{\text{Matching}} = \underbrace{\frac{1}{N_1} \sum_{i=1}^n D_i Y_i}_{\text{Treated group average}} - \underbrace{\frac{1}{N_1} \sum_{i=1}^n D_i Y_{j(i)}}_{\text{Matched control group average}}$$

## Matching estimator

$$\hat{\mathbb{E}}[Y_i(0) \mid D_i = 1] = \frac{1}{N_1} \sum_{i=1}^n D_i Y_{j(i)}$$

Each control unit has (approximately) the same  $\mathbf{X}$  as its respective treated unit so their  $Y(0)$  is a good estimate (on average) for the treated unit's  $Y(0)$ .

## Matching estimator

$$\hat{\mathbb{E}}[Y_i(0) \mid D_i = 1] = \frac{1}{N_1} \sum_{i=1}^n D_i Y_{j(i)}$$

Each control unit has (approximately) the same  $\mathbf{X}$  as its respective treated unit so their  $Y(0)$  is a good estimate (on average) for the treated unit's  $Y(0)$ .

The matched control units  $\{j(i) : D_i = 1\}$  have the same distribution of  $\mathbf{X}$  as the treated units (by design), so in some sense you are removing selection on  $\mathbf{X}$

## Defining “close”

How do we decide which control units are close? We need some measure of “distance”.

The simplest measure is the **euclidean distance**. For units  $i$  and  $j$ , the distance between  $X_i$  and  $X_j$ :

$$\sqrt{\sum_{k=1}^K (X_{i,k} - X_{j,k})^2}$$

## Defining “close”

How do we decide which control units are close? We need some measure of “distance”.

The simplest measure is the **euclidean distance**. For units  $i$  and  $j$ , the distance between  $\mathbf{X}_i$  and  $\mathbf{X}_j$ :

$$\sqrt{\sum_{k=1}^K (X_{i,k} - X_{j,k})^2}$$

In matrix notation this is

$$(\mathbf{X}_i - \mathbf{X}_j)'(\mathbf{X}_i - \mathbf{X}_j)$$

# Distances

$$\sqrt{\sum_{k=1}^K (X_{i,k} - X_{j,k})^2}$$

Issues with this distance metric:

- Puts more weight on covariates with larger variances (not scale invariant)
- Each covariate is viewed as being “equally important to match on”
- Highly collinear variables get “counted twice” in that their discrepancies get summed over twice

# Mahalanobis Distance

Can “scale” the distances using the variance-covariance matrix of  $\mathbf{X}_i$ :

$$S_{\mathbf{XX}'} \equiv \left[ 1/N \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \right]$$

- The  $(k, \ell)$ -th element is the sample covariance  $\text{Cov}(X_{i,k}, X_{i,\ell})$  and the diagonal being the variance of  $X_{i,k}$
- If the covariance between two elements is high, then we don’t want to give “credit” to both as “distinct” variables

## Mahalanobis Distance

The **Mahalanobis Distance** is the distance between unit  $i$  and  $j$ 's covariate rescaled by the inverse variance-covariance matrix:

$$(\mathbf{X}_i - \mathbf{X}_j)' S_{\mathbf{XX}'}^{-1} (\mathbf{X}_i - \mathbf{X}_j)$$

- All elements have variance 1 from the rescaling, so this measure is scale-invariant
- Covariates that are highly collinear do not get counted twice

## Ensuring we have matches

How do we know that there is going to be matchable control units?

- Perhaps units with a certain value (and nearby values) of  $X_i = x$  are all treated, so there are no (high-quality) controls to match to

## Ensuring we have matches

How do we know that there is going to be matchable control units?

- Perhaps units with a certain value (and nearby values) of  $X_i = x$  are all treated, so there are no (high-quality) controls to match to

Option 1: We require the **overlap** condition. There is an  $\epsilon > 0$  such that for all  $x \in \mathbb{X}$

$$1 - \epsilon > \mathbb{P}(D = 1 \mid \mathbf{X} = x) > \epsilon$$

- Guarantees that in large samples we will see some treated and control units for each value of  $\mathbf{X}$

## Ensuring we have matches

Option 2: Subset to treated units that have matchable units

- Changes our definition of the “population” and hence our ATT
- A more subtle point is that if we determine the subsetting based by on our finite sample, then inference gets more complicated

## Ensure 'balance'

It is very common to assess 'balance' between covariates after performing matching.

- Let's the reader verify that the matching is of reasonably high-quality (balance on included  $X$ s)
  - As a reader, don't give too much credit to this (they are trying to match them, so they should be close)
- Perhaps include other variables that were not included to suggest the treated and matched control group look similar

## Ensure 'balance'

You might be tempted to report a t-test on the difference in means of different  $X$  variables between your treated and matched sample

- As your sample size gets larger, you will find significant differences even if they are not *meaningful* differences

Imbens and Wooldridge (2009, JEL) recommend (as is commonly done) standardized difference-in-means:

$$\frac{\bar{X}_1 - \bar{X}_0}{\sqrt{S_1^2 + S_0^2}}$$

- $S_1^2$  and  $S_0^2$  are the sample variance of  $X$  for the treated and control-group respectively

## Example: Balance Improvements in the Lalonde Data

Covariate	CPS Controls (15992)		NSW Treated (185)		Mahalanobis Matched Controls (370)
	mean	(s.d.)	mean	(s.d.)	normalized diff.
Age	33.23	(11.05)	25.82	(7.16)	-0.16
Education	12.03	(2.87)	10.35	(2.01)	-0.09
Married	0.71	(0.45)	0.19	(0.39)	-0.20
Nodegree	0.30	(0.46)	0.71	(0.46)	0.18
Black	0.07	(0.26)	0.84	(0.36)	0.00
Hispanic	0.07	(0.26)	0.06	(0.24)	0.00
Earn '74	14.02	(9.57)	2.10	(4.89)	0.00
Earn '74 positive	0.88	(0.32)	0.29	(0.46)	0.00
Earn '75	13.65	(9.27)	1.53	(3.22)	-0.01
Earn '75 positive	0.89	(0.31)	0.40	(0.49)	0.00

# Inference after Matching

Note that matching takes on two steps:

1. Create a matched sample using information on  $X_i$
2. Estimate a difference-in-means (using regression)

The standard errors on the regression estimate ignore the noise generated in step 1

- Incorporating uncertainty from step 1 is a very difficult problem (Guido Imbens and Alberto Abadie, *Econometrica* 2006)

## Coding implementation

In R, use the MatchIt package

In Stata, use the teffects nnmatch command

# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Reminder of Setup

We observe a sample of  $n$  observations,  $Y$  is outcome,  $\mathbf{X}$  is a vector of control variables,  $D$  is our treatment variable of interest. We assume the conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

## Reminder of Setup

We observe a sample of  $n$  observations,  $Y$  is outcome,  $\mathbf{X}$  is a vector of control variables,  $D$  is our treatment variable of interest. We assume the conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

Our average-treatment effect estimate can be written as the average of CATE (averaged over distribution of  $X$ )

$$\tau_{ATE} = \mathbb{E}[ \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] ]$$

# Propensity Score

Our previous estimator relied on proper modeling of the conditional expectation function. This method, instead, will rely on estimation of the **propensity score** (Rosenbaum and Rubin, 1983):

$$\pi(\mathbf{x}) = \mathbb{P}(D_i = 1 \mid \mathbf{X}_i = \mathbf{x})$$

- Looking at units with  $\mathbf{X}_i = \mathbf{x}$ , the propensity score is the probability a random unit in this subset is treated
- This was discussed when defining **overlap**:  $0 < \pi(\mathbf{x}) < 1$

# Properties of Propensity score

Property 1:  $\mathbf{X}$  and  $D_i$  are independent conditional on the propensity-score

$$\mathbf{X} \perp\!\!\!\perp D_i \mid \pi(\mathbf{X})$$

- Treated and control units with the same  $\pi(X_i) = p$  have the same distribution of  $\mathbf{X}$
- Can assess this by binning the propensity score (into intervals  $\underline{p}, \bar{p}$ ) and checking balance

# Properties of Propensity score

Property 2: The original conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

implies that  $Y(0)/Y(1)$  and  $D$  are independent conditional on the propensity score

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \pi(\mathbf{X}_i)$$

## Propensity score CIA

This means instead of trying to match on a high-dimensional vector  $\mathbf{X}$ , we could instead match on the scalar  $\pi(\mathbf{X}_i)$

- We in essence have collapsed all the relevant information about  $\mathbf{X}$  down to the propensity-score
- However, now our task is estimating  $\pi(\mathbf{x}) = \mathbb{P}(D_i = 1 \mid \mathbf{X}_i = \mathbf{x})$

## Estimation of Propensity score

Since our outcome variable,  $D_i$  is binary, it is most common to estimate the propensity score as a logistic function:

$$\pi(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\gamma})}{1 + \exp(\mathbf{x}'\boldsymbol{\gamma})}$$

- This forces fitted values to be between zero and 1
- Again, we want to have a *flexible* model (polynomials, interactions, etc.)

## Estimation of Propensity score

Since our outcome variable,  $D_i$  is binary, it is most common to estimate the propensity score as a logistic function:

$$\pi(\mathbf{x}) = \frac{\exp(\mathbf{x}'\boldsymbol{\gamma})}{1 + \exp(\mathbf{x}'\boldsymbol{\gamma})}$$

- This forces fitted values to be between zero and 1
- Again, we want to have a *flexible* model (polynomials, interactions, etc.)

Can estimate using `feglm` with `family = binomial(link = "logit")` logit in Stata

## Matching on propensity-score method

We can proceed as before with our matching method using the propensity-score as the only covariate. This means we match each treated unit to the control unit(s) with the closest estimated propensity score  $\hat{\pi}_i$

Then perform difference-in-means estimator on matched control group

## Overlap condition

This estimator also relies on the overlap condition: there is an  $\epsilon > 0$  such that for all  $x \in \mathbb{X}$

$$1 - \epsilon > \mathbb{P}(D = 1 \mid \mathbf{X} = x) > \epsilon$$

- When matching on the propensity score, we also want the control group to have propensity scores along the full range of treated unit's propensity scores

# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

**Inverse probability of treatment weighting**

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Reminder of Setup

We assume the conditional independence assumption

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid \mathbf{X}_i = \mathbf{x}$$

We have the propensity score and overlap condition. For all  $\mathbf{x}$ ,

$$0 < \pi(\mathbf{x}) \equiv \mathbb{P}(D_i = 1 \mid \mathbf{X}_i = \mathbf{x}) < 1$$

Our estimand of interest is the ATE/ATT

$$\tau_{ATE} = \mathbb{E}[ \mathbb{E}[Y_i(1) \mid \mathbf{X}_i = \mathbf{x}] - \mathbb{E}[Y_i(0) \mid \mathbf{X}_i = \mathbf{x}] ]$$

# Difference-in-means Estimator

Our difference-in-means estimator can be written as

$$\hat{\tau}_{\text{DIM}} = \mathbb{E}_n \left[ \frac{W_i}{n_1} Y_i \right] - \mathbb{E}_n \left[ \frac{1 - W_i}{n_0} Y_i \right]$$

- All control units are weighted the same, regardless of how “similar” they are to the treated unit

## Inverse Probability of Treatment

Instead, consider weighing by the inverse probability of treatment:

$$\mathbb{E}\left[\frac{(1 - W_i)Y_i}{1 - \pi(\mathbf{X}_i)}\right]$$

# Inverse Probability of Treatment

Instead, consider weighing by the inverse probability of treatment:

$$\mathbb{E}\left[\frac{(1 - W_i)Y_i}{1 - \pi(\mathbf{X}_i)}\right] = \mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0)\right]$$

# Inverse Probability of Treatment

Instead, consider weighing by the inverse probability of treatment:

$$\begin{aligned}\mathbb{E}\left[\frac{(1 - W_i)Y_i}{1 - \pi(\mathbf{X}_i)}\right] &= \mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0) \mid \mathbf{X}_i\right]\right]\end{aligned}$$

# Inverse Probability of Treatment

Instead, consider weighing by the inverse probability of treatment:

$$\begin{aligned}\mathbb{E}\left[\frac{(1 - W_i)Y_i}{1 - \pi(\mathbf{X}_i)}\right] &= \mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0) \mid \mathbf{X}_i\right]\right] \\ &= \mathbb{E}\left[\frac{(1 - \pi(\mathbf{X}_i))}{1 - \pi(\mathbf{X}_i)} \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right]\end{aligned}$$

# Inverse Probability of Treatment

Instead, consider weighing by the inverse probability of treatment:

$$\begin{aligned}\mathbb{E}\left[\frac{(1 - W_i)Y_i}{1 - \pi(\mathbf{X}_i)}\right] &= \mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)}Y_i(0) \mid \mathbf{X}_i\right]\right] \\ &= \mathbb{E}\left[\frac{(1 - \pi(\mathbf{X}_i))}{1 - \pi(\mathbf{X}_i)} \mathbb{E}[Y_i(0) \mid \mathbf{X}_i]\right] \\ &= \mathbb{E}[\mathbb{E}[Y_i(0) \mid \mathbf{X}_i]] = \mathbb{E}[Y_i(0)],\end{aligned}$$

where the second equality is the LIE and the third is the definition of the propensity score

## Weighting Control Units

$$\mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i\right] = \mathbb{E}[Y_i(0)]$$

The IPTW estimator puts weight  $\frac{1}{1 - \pi(\mathbf{X}_i)}$  on control unit  $i$

- The weights get larger for units with *large* values of  $\pi(\mathbf{X}_i)$ , i.e. the units that we think really should be treated but are not

# Inverse Probability of Treatment

## *ATE Estimator*

Using similar math for  $\frac{W_i Y_i}{\pi_i(\mathbf{X})}$ , we have the Horvitz-Thompson (1952) “inverse probability weighting” estimator

$$\begin{aligned}\hat{\tau}_{\text{IPTW,ATE}} &\equiv \mathbb{E}\left[\frac{W_i}{\pi(\mathbf{X}_i)} Y_i\right] - \mathbb{E}\left[\frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i\right] \\ &= \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]\end{aligned}$$

# Inverse Probability of Treatment

## *ATT Estimator*

For the ATT estimator, we can just average  $Y_i$  for the treated group and modify our control group weights

$$\begin{aligned}\hat{\tau}_{\text{IPTW,ATT}} &\equiv \mathbb{E}\left[\frac{W_i}{n_1} Y_i\right] - \mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i\right] \\ &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 1]\end{aligned}$$

## ATT proof

$$\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i \mid \mathbf{X}_i\right]\right]$$

## ATT proof

$$\begin{aligned}\mathbb{E} \left[ \frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i \right] &= \mathbb{E} \left[ \mathbb{E} \left[ \frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i \mid \mathbf{X}_i \right] \right] \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \frac{\pi(\mathbf{X}_i)}{n_1} Y_i(0) \mid \mathbf{X}_i \right] \right] \\ &= \mathbb{E} \left[ \mathbb{E} \left[ \frac{D_i}{n_1} Y_i(0) \mid \mathbf{X}_i \right] \right]\end{aligned}$$

## ATT proof

$$\begin{aligned}\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} \frac{(1 - W_i)}{1 - \pi(\mathbf{X}_i)} Y_i \mid \mathbf{X}_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{\pi(\mathbf{X}_i)}{n_1} Y_i(0) \mid \mathbf{X}_i\right]\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[\frac{D_i}{n_1} Y_i(0) \mid \mathbf{X}_i\right]\right] \\ &= \mathbb{E}[\mathbb{E}[Y_i(0) \mid \mathbf{X}_i, D_i = 1]] \\ &= \mathbb{E}[Y_i(0) \mid D_i = 1]\end{aligned}$$

# Overlap

Note that our weights divide by  $\pi(\mathbf{X}_i)$  or  $1 - \pi(\mathbf{X}_i)$

- Overlap is necessary for all  $\mathbf{X}_i$  so that we are not dividing by 0!

# Overlap

Note that our weights divide by  $\pi(\mathbf{X}_i)$  or  $1 - \pi(\mathbf{X}_i)$

- Overlap is necessary for all  $\mathbf{X}_i$  so that we are not dividing by 0!

In settings where  $\pi(\mathbf{X}_i)$  is very close to 0 or 1, those observations are given *HUGE* weights

- $\pi(\mathbf{X}_i) = 0.0001$  implies a treatment weight of  $1/0.0001 = 10000$

# Overlap

Note that our weights divide by  $\pi(\mathbf{X}_i)$  or  $1 - \pi(\mathbf{X}_i)$

- Overlap is necessary for all  $\mathbf{X}_i$  so that we are not dividing by 0!

In settings where  $\pi(\mathbf{X}_i)$  is very close to 0 or 1, those observations are given *HUGE* weights

- $\pi(\mathbf{X}_i) = 0.0001$  implies a treatment weight of  $1/0.0001 = 10000$   
⇒ very noisy estimates

## Trimming the propensity scores

It is common to “trim” (or drop) data with propensity scores with really large or really small propensity scores (see discussion in Imbens and Xu (2024))

- Dehejia and Wahba (1999) recommend dropping control units with propensity scores smaller than the smallest treated unit’s propensity score
- Crump et. al. (2009) recommend trimming data to  $\hat{\pi}(\mathbf{X}_i) \in [0.1, 0.9]$

## Hájek weights

Our IPTW weights in the Horvitz-Thompson estimator were given by

$$w_{1,i} = \frac{D_i}{\pi(\mathbf{X}_i)} \text{ and } w_{0,i} = \frac{1 - D_i}{1 - \pi(\mathbf{X}_i)}$$

- These weights do not sum to 1 which makes these estimates extra noisy

## Hájek weights

Our IPTW weights in the Horvitz-Thompson estimator were given by

$$w_{1,i} = \frac{D_i}{\pi(\mathbf{X}_i)} \text{ and } w_{0,i} = \frac{1 - D_i}{1 - \pi(\mathbf{X}_i)}$$

- These weights do not sum to 1 which makes these estimates extra noisy

It is more efficient (Hirano, Imbens, Ridder 2003) to normalize these weights to sum to 1.

This is called the Hájek estimator:

$$\tilde{w}_{1,i} = \frac{w_{1,i}}{1/n \sum_i w_{1,i}} \text{ and } \tilde{w}_{0,i} = \frac{w_{0,i}}{1/n \sum_i w_{0,i}}$$

## Estimation of propensity score

Since  $D_i$  is an indicator variable, we (may) want to avoid a linear regression

- Can produce estimates that are outside of  $(0, 1)$  creating problems for our estimator

## Estimation of propensity score

Since  $D_i$  is an indicator variable, we (may) want to avoid a linear regression

- Can produce estimates that are outside of  $(0, 1)$  creating problems for our estimator

Instead, use a logistic regression via `fixest` package and `feglm(family = "logit")`:

$$\text{logit}(Y_i) = X_i\beta + u_i$$

## Coding implementation

In R, use the `WeightIt` package to estimate the propensity score weights and the IPTW estimators

In Stata, use the `teffects ipw` command

# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

## Simplest Estimator under Conditional Independence

Our work so far has started from the definition of the treatment effect of interest and worked with that definition to inform us of an estimator

Applied researchers seem to work the other way:

- Run a regression and ask what it identifies

# Simplest Estimator under Conditional Independence

The simplest regression you would think to do is

$$Y_i = \tau D_i + \mathbf{X}'_i \beta + u_i$$

- Assuming  $\mathbf{X}_i$  contains an intercept
- $\hat{\tau}_{OLS}$  serves as our estimate of  $\tau_{ATE}$

Is  $\hat{\tau}_{OLS}$  a good estimate?

## What does $\hat{\tau}_{OLS}$ estimate?

We know from the FWL theorem that our OLS estimate is

$$\begin{aligned}\hat{\tau}_{OLS} &= \frac{\text{Cov}(\tilde{D}_i, \tilde{Y}_i)}{\text{Var}(\tilde{D}_i)} = \mathbb{E}\left[\frac{\tilde{D}_i}{\text{Var}(\tilde{D}_i)} \tilde{Y}_i\right] \\ &= \mathbb{E}\left[\frac{\tilde{D}_i}{\text{Var}(\tilde{D}_i)} Y_i\right]\end{aligned}$$

## What does $\hat{\tau}_{OLS}$ estimate?

We know from the FWL theorem that our OLS estimate is

$$\begin{aligned}\hat{\tau}_{OLS} &= \frac{\text{Cov}(\tilde{D}_i, \tilde{Y}_i)}{\text{Var}(\tilde{D}_i)} = \mathbb{E}\left[\frac{\tilde{D}_i}{\text{Var}(\tilde{D}_i)} \tilde{Y}_i\right] \\ &= \mathbb{E}\left[\frac{\tilde{D}_i}{\text{Var}(\tilde{D}_i)} Y_i\right]\end{aligned}$$

Our estimator is a weighted average of  $Y_i$  with weight  $\tilde{D}_i / \text{Var}(\tilde{D}_i)$

- $\tilde{D}_i$  is large if the treatment status is very different than what we predicted it to be using our linear model

## What does $\hat{\tau}_{0\text{LS}}$ estimate?

If we assume that treatment effects do not vary across individuals, what we call **homogeneous effects**, then this weighting does not matter and

$$\hat{\tau}_{0\text{LS}} \xrightarrow{p} \tau_{\text{ATE}}$$

## What does $\hat{\tau}_{OLS}$ estimate?

If we assume that treatment effects do not vary across individuals, what we call **homogeneous effects**, then this weighting does not matter and

$$\hat{\tau}_{OLS} \xrightarrow{p} \tau_{ATE}$$

- Homogeneous effects are unlikely to hold in many settings where the effects of treatment may vary based on characteristics of the individual  $X_i$

# Heterogeneous effects

When effects are heterogeneous, things get weird

- We'll talk about it now, because you will see a lot of linear regression in life

# Heterogeneous effects

When effects are heterogeneous, things get weird

- We'll talk about it now, because you will see a lot of linear regression in life

Aronow and Samii (2016, AJPS) work with this OLS estimator and makes very favorable assumptions (linearity is correct):

1. Conditional independence assumption
2.  $Y_i(0) = \mathbf{X}'_i \beta + u_i$ , i.e. outcomes are linear in  $\mathbf{X}_i$
3.  $\pi(\mathbf{X}_i) = \mathbf{X}'_i \gamma$ , i.e. propensity-score is linear in  $\mathbf{X}_i$

## Aronow and Samii (2016, AJPS)

Under these three conditions, you can show  $\hat{\tau}_{OLS} \xrightarrow{p} \mathbb{E}\left[\frac{w_i}{\mathbb{E}[w_i]}\tau_i\right]$

- $\tau_i = Y_i(1) - Y_i(0)$  and  $\frac{w_i}{\mathbb{E}[w_i]} \geq 0$  is the weight put on that unit's treatment effect

Our estimates are some weighted average of treatment effects. Not the ATE but some weighted average, at least

- Note this is the most favorable conditions for this regression (besides homogeneous treatment effects)

## Who gets a lot of weight

In fact, the weight given to units is given by  $w_i = (D_i - \mathbb{E}[D_i | \mathbf{X}_i])^2$ , so units whose treatment status were unexpected get more weight!

## Who gets a lot of weight

In fact, the weight given to units is given by  $w_i = (D_i - \mathbb{E}[D_i | \mathbf{X}_i])^2$ , so units whose treatment status were unexpected get more weight!

E.g. from Aronow and Samii look at cross-country treatment effect estimate:



# Roadmap

Conditional Independence Assumption

Regression Adjustment

Matching estimators

Propensity-score matching

Inverse probability of treatment weighting

Why not Linear Regression?

More ‘flexible’ approaches and double-robustness

# Regression Adjustment and Propensity-scores

We have seen two key methods:

1. One relies on estimation of the conditional expectation functions

$$\mu_d(\mathbf{x}) = \mathbb{E}[Y_i | \mathbf{X}_i = \mathbf{x}, D_i = d]$$

2. The other relies on estimation of the propensity score  $\pi(\mathbf{x}) = \mathbb{P}(D_i = 1 | \mathbf{X}_i = \mathbf{x})$

Both will suffer problems if you specify an incorrect functional form (e.g. you incorrectly assume linearity in  $\mathbf{X}_i$ )

- This is true even if we have that conditional independence assumption and overlap holding!

# Approach 1

## *Machine Learning*

One thing that folks can do is leverage machine learning methods to more flexibly model  $\mu_d(\mathbf{x})$  and  $\pi(\mathbf{x})$

- Make the models very flexible (polynomials, bins, interactions between  $\mathbf{X}$ , etc.)

However, this runs the risk of over-fitting and creating very noisy (and hence unreliable) estimates

# Approach 1

## *Machine Learning*

One thing that folks can do is leverage machine learning methods to more flexibly model  $\mu_d(\mathbf{x})$  and  $\pi(\mathbf{x})$

- Make the models very flexible (polynomials, bins, interactions between  $\mathbf{X}$ , etc.)

However, this runs the risk of over-fitting and creating very noisy (and hence unreliable) estimates

Machine learning methods (and cross-validation) can be used to prevent overfitting while remaining flexible

- See “Double/Debiased Machine Learning” paper in Econometrics Journal for approachable introduction

## Approach 2

### *Doubly-robust estimators*

A more traditional approach is to use a **doubly-robust estimator** where you use estimates for both  $\mu_d(x)$  and  $\pi(x)$

- The idea is to have “2 shots” at getting it right

# Doubly-robust treatment effect estimator

The augmented IPW (AIPW) estimator of Robins, Rotnitzky, and Zhao (1994):

$$\begin{aligned}\hat{\tau}_{\text{DR}} = & \frac{1}{n} \sum_{i=1}^n (\hat{\mu}_1(\mathbf{X}_i) - \hat{\mu}_1(\mathbf{X}_i)) \\ & + \frac{1}{n} \sum_{i=1}^n \left( \frac{D_i}{\hat{\pi}(\mathbf{X}_i)} (Y_i - \hat{\mu}_1(\mathbf{X}_i)) - \frac{(1 - D_i)}{1 - \hat{\pi}(\mathbf{X}_i)} (Y_i - \hat{\mu}_0(\mathbf{X}_i)) \right)\end{aligned}$$

- A regression adjustment estimator + an adjustment term

## Adjustment term

$$\frac{1}{n} \sum_{i=1}^n \left( \frac{D_i}{\hat{\pi}(\mathbf{X}_i)} (Y_i - \hat{\mu}_1(\mathbf{X}_i)) - \frac{(1 - D_i)}{1 - \hat{\pi}(\mathbf{X}_i)} (Y_i - \hat{\mu}_0(\mathbf{X}_i)) \right)$$

The adjustment term  $(Y_i - \hat{\mu}_d(\mathbf{X}_i))$  is the prediction error of our model

- This includes the part of  $\mu_d(\mathbf{X}_i)$  that we failed to estimate
- The remaining part biases our treatment effect

Take the prediction error and perform IPW on it to estimate the remaining treatment effect

## “Doubly-robust”

The AIPW estimator has two great benefits:

1. It is consistent so long as either  $\mu_d(x)$  or  $\pi(x)$  are correctly specified (only 1 is needed, but both is better!)

## “Doubly-robust”

The AIPW estimator has two great benefits:

1. It is consistent so long as either  $\mu_d(x)$  or  $\pi(x)$  are correctly specified (only 1 is needed, but both is better!)
2. If both are correct, then this estimator obtains the ‘semi-parametric efficiency bound’ (as efficient as can be)

## “Doubly-robust”

The AIPW estimator has two great benefits:

1. It is consistent so long as either  $\mu_d(x)$  or  $\pi(x)$  are correctly specified (only 1 is needed, but both is better!)
2. If both are correct, then this estimator obtains the ‘semi-parametric efficiency bound’ (as efficient as can be)

While theoretically this has great qualities, it is more data hungry

- In smaller finite samples, these estimators can be noisy