

Lecture 07:
Selection on observables

PPHA 34600
Prof. Fiona Burlig


Harris School of Public Policy
University of Chicago

From last time: be cautious with non-experimental findings

We looked at a LaLonde-style evaluation of non-RCT methods:

- These approaches did not nail the experimental result
- The more opportunity for selection, the worse they did

Why did the non-experimental estimators fail?

The root of all  in this class is selection:


① Selection on observables

- Are treated and untreated units different in ways we can observe?

② Selection on unobservables

- Are treated and untreated units different in ways we can't observe?

Why did the non-experimental estimators fail?

The root of all  in this class is selection:

- 1 Selection on observables
 - Are treated and untreated units different in ways we can observe?
- 2 Selection on unobservables
 - Are treated and untreated units different in ways we can't observe?

Suppose we're no longer in a randomized world

We still want to estimate treatment effects

- Our original instinct was the naive estimator: $\bar{Y}(1) - \bar{Y}(0)$
- Assumes *all* differences between $D_i = 1$ and $D_i = 0$ are “as good as random”
- We can weaken this assumption if we see other characteristics
- We will turn to a series of designs where we “control for stuff”

We are entering the world of **selection on observables** designs:
We will assume that, conditional on observables, treatment assignment is independent of potential outcomes (💀?)

Selection on observables is a form of last-resort design:

- This section should feel extremely unsatisfying
- That is on purpose!
- These designs are typically not (very) believable

Central assumption underlying SOO designs

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

In words:

- Potential outcomes are independent of D_i , conditional on covariates

Central assumption underlying SOO designs

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

In words:

- Potential outcomes are independent of D_i , conditional on covariates

In other words:

- “Conditional unconfoundedness”

In different words:

- “Conditional independence”

In other different words:

- “Strongly ignorable treatment assignment”

In other different words:

- Once we control for X_i , treatment is as good as random

In the last set of words:

- Once we control for X_i , we've eliminated selection

We actually need a second assumption too

$$0 < Pr(D_i = 1 | X_i = x) < 1$$

In words:

- The probability that $D_i = 1$ for all levels of X_i is between zero and one

We actually need a second assumption too

$$0 < Pr(D_i = 1|X_i = x) < 1$$

In words:

- The probability that $D_i = 1$ for all levels of X_i is between zero and one

In other words:

- “Common support”

In different words:

- There are both treated and untreated units for each level of X

In other different words:

- “Overlap”

What do these assumptions buy us?

Recall that we're trying to estimate the ATE:

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

...but all we can actually see is $E[Y_i(1)|D_i = 1]$ and $E[Y_i(0)|D_i = 0]$

What do these assumptions buy us?

Recall that we're trying to estimate the ATE:

$$\tau^{ATE} = E[Y_i(1)] - E[Y_i(0)]$$

...but all we can actually see is $E[Y_i(1)|D_i = 1]$ and $E[Y_i(0)|D_i = 0]$

Under random assignment, we had that

$$(Y_i(1), Y_i(0)) \perp D_i$$

This implies that:

$$E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$$

and

$$E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$$

so we could just estimate

$$\tau^{ATE} = \bar{Y}(1) - \bar{Y}(0)$$

What do these assumptions buy us?

Our new assumption says something a bit weaker:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

What do these assumptions buy us?

Our new assumption says something a bit weaker:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

$$\begin{aligned} E[Y_i(1) | D_i = 1, X_i = x] &= E[Y_i(1) | D_i = 0, X_i = x] \\ &= E[Y_i(1) | D_i, X_i = x] = E[Y_i(1) | X_i = x] \end{aligned}$$

and

$$\begin{aligned} E[Y_i(0) | D_i = 1, X_i = x] &= E[Y_i(0) | D_i = 0, X_i = x] \\ &= E[Y_i(0) | D_i, X_i = x] = E[Y_i(0) | X_i = x] \end{aligned}$$

What do these assumptions buy us?

Our new assumption says something a bit weaker:

$$(Y_i(1), Y_i(0)) \perp D_i | X_i$$

$$\begin{aligned} E[Y_i(1) | D_i = 1, X_i = x] &= E[Y_i(1) | D_i = 0, X_i = x] \\ &= E[Y_i(1) | D_i, X_i = x] = E[Y_i(1) | X_i = x] \end{aligned}$$

and

$$\begin{aligned} E[Y_i(0) | D_i = 1, X_i = x] &= E[Y_i(0) | D_i = 0, X_i = x] \\ &= E[Y_i(0) | D_i, X_i = x] = E[Y_i(0) | X_i = x] \end{aligned}$$

So we can write:

$$\tau^{SOO} = E[Y_i(1) | X_i = x] - E[Y_i(0) | X_i = x]$$

Estimating τ^{ATE} under SOO

$$\tau^{SOO} = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

Now integrate across all values of X_i (take a weighted average):

$$\int (E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x])dP(X)$$

Estimating τ^{ATE} under SOO

$$\tau^{SOO} = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

Now integrate across all values of X_i (take a weighted average):

$$\int (E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x])dP(X)$$

$$= \underbrace{E[Y_i(1)] - E[Y_i(0)]}_{\text{by calculus } \text{💀}}$$

Estimating τ^{ATE} under SOO

$$\tau^{SOO} = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

Now integrate across all values of X_i (take a weighted average):

$$\int (E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x])dP(X)$$

$$= \underbrace{E[Y_i(1)] - E[Y_i(0)]}$$

by calculus 🦴

$$= \underbrace{\tau^{ATE}}$$

by definition

Estimating τ^{ATE} under SOO

$$\tau^{SOO} = E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]$$

Now integrate across all values of X_i (take a weighted average):

$$\int (E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x])dP(X)$$

$$= \underbrace{E[Y_i(1)] - E[Y_i(0)]}$$

by calculus ☠

$$= \underbrace{\tau^{ATE}}$$

by definition

Under conditional independence and common support, we can get from τ^{SOO} to τ^{ATE} !

How do we actually estimate τ^{SOO} ?

There are two main SOO designs:

- 1 Regression adjustment
 - Controlling for stuff
- 2 Matching
 - Pairing treated and untreated on observables

These are all fancy ways to estimate

$$\tau^{ATE} = \int E[Y_i(1)|X_i = x] - E[Y_i(0)|X_i = x]dP(X)$$

Approach 1: Regression adjustment

Our regression model back in potential outcomes land:

$$Y_i(0) = \alpha + \gamma X_i + \nu_i$$

$$Y_i(1) = Y_i(0) + \tau + \tau_i$$

Approach 1: Regression adjustment

Our regression model back in potential outcomes land:

$$Y_i(0) = \alpha + \gamma X_i + \nu_i$$

$$Y_i(1) = Y_i(0) + \tau + \tau_i$$

Under constant treatment effects:

$$Y_i(1) = Y_i(0) + \tau + \underbrace{0}_{\text{no } i \text{ specific bit}}$$

Approach 1: Regression adjustment

Our regression model back in potential outcomes land:

$$Y_i(0) = \alpha + \gamma X_i + \nu_i$$

$$Y_i(1) = Y_i(0) + \tau + \tau_i$$

Under constant treatment effects:

$$Y_i(1) = Y_i(0) + \tau + \underbrace{0}_{\text{no } i \text{ specific bit}}$$

We can just write this as:

$$Y_i = \alpha + \tau D_i + \gamma X_i + \nu_i$$

Note that we're used to just working with

$$\varepsilon_i = \gamma X_i + \nu_i$$

In randomized land, everything works nicely

Under random assignment, we have that:

- $Y_i \perp D_i$
- AKA $E[\varepsilon_i | D_i] = 0$
- AKA $E[(\gamma X_i + \nu_i) | D_i] = 0$

This lets us estimate:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

and have $\hat{\tau} \approx \tau^{ATE}$

→ Note that by randomization, we don't have to worry about the X_i s!

Regression with selection on observables

Under selection on observables, we have that:

- $Y_i \perp D_i | X_i$
- AKA $E[\varepsilon_i | D_i, X_i] = 0$
- AKA $E[(\gamma X_i + \nu_i) | D_i, X_i] = 0$

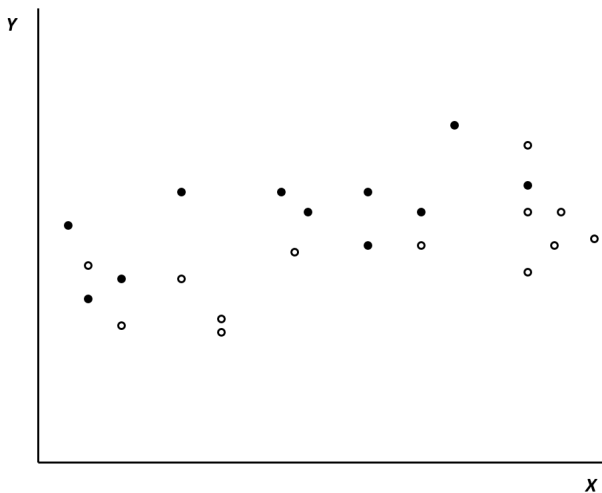
Now we have to estimate:

$$Y_i = \alpha + \tau D_i + \gamma X_i + \nu_i$$

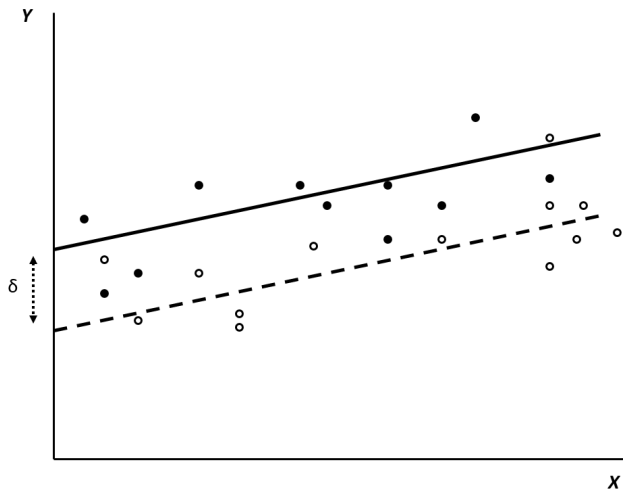
to get $\hat{\tau} \approx \tau^{ATE}$

→ Now we have *conditional* independence: if we leave X_i out, we're in trouble, because $E[\varepsilon_i | D_i]$ is not necessarily zero anymore!

Selection on observables: OLS



Selection on observables: OLS



Concerns with regression adjustment

When we run

$$Y_i = \alpha + \tau D_i + \gamma X_i + \nu_i$$

We can represent this as a difference in means between treated and untreated units:

$$\bar{Y}_U = \alpha + \gamma \bar{X}_U$$

and

$$\bar{Y}_T = \alpha + \tau + \gamma \bar{X}_T$$

Concerns with regression adjustment

When we run

$$Y_i = \alpha + \tau D_i + \gamma X_i + \nu_i$$

We can represent this as a difference in means between treated and untreated units:

$$\bar{Y}_U = \alpha + \gamma \bar{X}_U$$

and

$$\bar{Y}_T = \alpha + \tau + \gamma \bar{X}_T$$

$$\underbrace{(\bar{Y}_T - \bar{Y}_U)}_{\text{subtraction}} = \tau + \gamma(\bar{X}_T - \bar{X}_U)$$

Concerns with regression adjustment

When we run

$$Y_i = \alpha + \tau D_i + \gamma X_i + \nu_i$$

We can represent this as a difference in means between treated and untreated units:

$$\bar{Y}_U = \alpha + \gamma \bar{X}_U$$

and

$$\bar{Y}_T = \alpha + \tau + \gamma \bar{X}_T$$

$$\underbrace{(\bar{Y}_T - \bar{Y}_U)}_{\text{subtraction}} = \tau + \gamma(\bar{X}_T - \bar{X}_U)$$

$$\hat{\tau} = \underbrace{(\bar{Y}_T - \bar{Y}_U) - \hat{\gamma}(\bar{X}_T - \bar{X}_U)}_{\text{rearranged}}$$

Functional form assumptions

We rely heavily on two things:

- 1 \bar{X}_T being close to \bar{X}_U
 - If $|\bar{X}_T - \bar{X}_U|$ is large, our estimate of $\hat{\tau}$ will be biased
 - We need “good overlap” between X_i for control and treatment
 - What does this mean when we have multiple X_i s?

Functional form assumptions

We rely heavily on two things:

- 1 \bar{X}_T being close to \bar{X}_U
 - If $|\bar{X}_T - \bar{X}_U|$ is large, our estimate of $\hat{\tau}$ will be biased
 - We need “good overlap” between X_i for control and treatment
 - What does this mean when we have multiple X_i s?
- 2 Our assumed functional form
 - Our regression assumes the true relationship is $Y_i = \alpha + \tau D_i + \gamma X_i$
 - We actually need to control for $E[D_i|X_i]$, not just X_i
 - We should have run: $Y_i = \alpha + \tau D_i + \gamma E[D_i|X_i] + \nu_i$
 - If $X_i \neq E[D_i|X_i]$, then $\gamma(X_i - E[D_i|X_i])$ is in our error term

→ $E[\nu_i|D_i, X_i] \neq 0$ ☠

Approach 2: Matching

We can avoid some concerns by matching:

- We compare untreated units to treated units with **identical** X_i s
- Difference in outcomes between treated and untreated is our $\hat{\tau}$

Approach 2: Matching

We can avoid some concerns by matching:

- We compare untreated units to treated units with **identical** X_i s
- Difference in outcomes between treated and untreated is our $\hat{\tau}$
- Since we're comparing identical X_i s:
 - We guarantee treated and control units have similar X_i

Approach 2: Matching

We can avoid some concerns by matching:

- We compare untreated units to treated units with **identical** X_i s
- Difference in outcomes between treated and untreated is our $\hat{\tau}$
- Since we're comparing identical X_i s:
 - We guarantee treated and control units have similar X_i
 - Functional form is irrelevant
- Still requires:
 - $Y_i \perp D_i | X_i$
 - $0 < Pr(D_i = 1 | X_i = x) < 1$

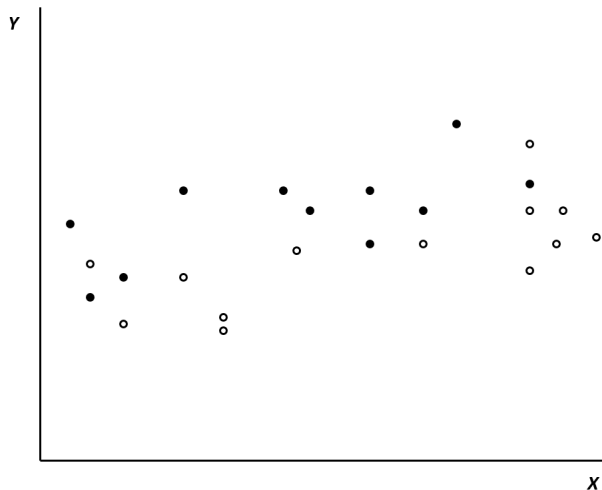
The exact matching estimator

The simplest possible matching estimator is exact matching:

- 1 Divide data into “cells” uniquely defined by the covariates
- 2 For each value of $X = x$ (each cell), calculate \bar{Y}_T and \bar{Y}_U
- 3 Calculate $\bar{Y}_T - \bar{Y}_U$ for each $X = x$
- 4 Estimate τ^{ATE} as a weighted average of (3)

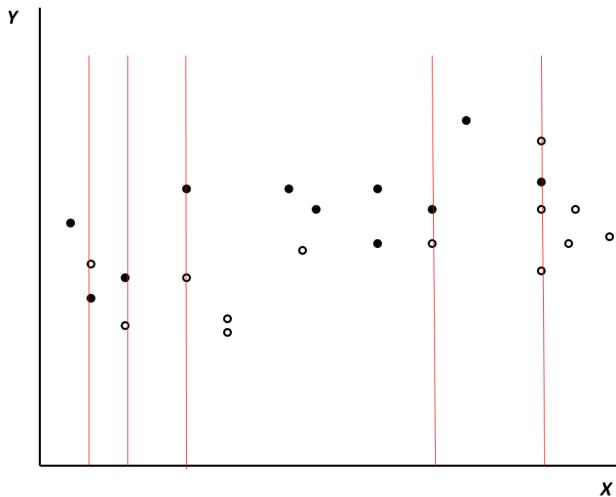
Note: This works for more than one X ! See additional slides.

Getting “close”



How would we implement exact matching?

Getting “close”



The Curse of Dimensionality

We're often interested in matching on multiple X s:

- You have to be very lucky (dumb?) to think selection on only one X !
- Much more likely: selection depends on many X s
- But the more X s you have, the less likely you are to have a match
- (This same issue bites for regression too)

The Curse of Dimensionality

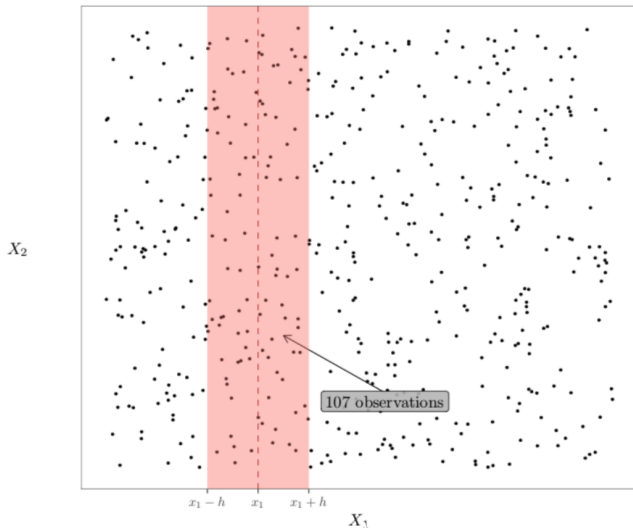
We're often interested in matching on multiple X s:

- You have to be very lucky (dumb?) to think selection on only one X !
- Much more likely: selection depends on many X s
- But the more X s you have, the less likely you are to have a match
- (This same issue bites for regression too)
- From *my* PhD econometrics class:

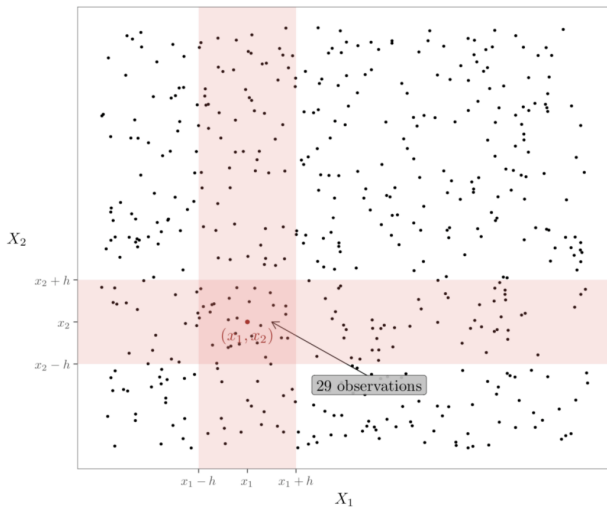
“Although you can sometimes reduce the dimensionality problems by making various parametric assumptions...you can never truly defeat the Curse of Dimensionality. It is, after all, a curse.”

– Michael L. Anderson, UC Berkeley

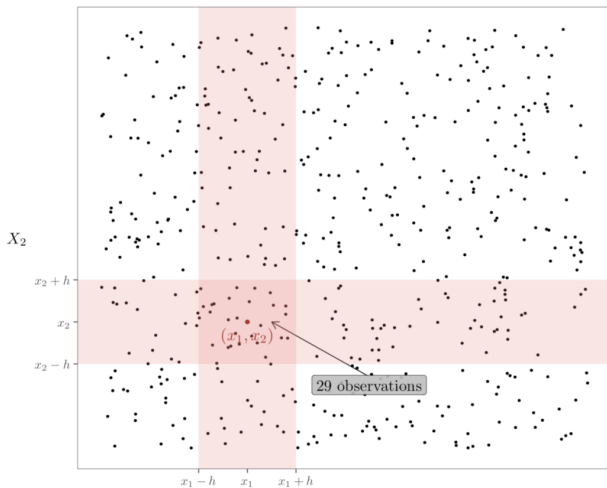
The Curse of Dimensionality



The Curse of Dimensionality



The Curse of Dimensionality



For $K = 30$ binary covariates, you'd need $N = 2^{30+1} = 2,147,483,648$



Examining the exact matching estimator

The good news:

- Creates observably identical treated and untreated comparisons
 - No need to worry about \bar{X}_T and \bar{X}_U being far apart
 - Makes no functional form assumptions
 - Don't have to worry about **how** to control for X s
- This is a very flexible estimator!

Examining the exact matching estimator

The good news:

- Creates observably identical treated and untreated comparisons
 - No need to worry about \bar{X}_T and \bar{X}_U being far apart
- Makes no functional form assumptions
 - Don't have to worry about **how** to control for X s

→ This is a very flexible estimator!

The bad news:

- It doesn't work for *continuous* X s!
 - How do you define cells of continuous variables?

→ Very flexible, but not super practically useful?

Going beyond the exact matching estimator


What can we do when we have continuous X ?

- For each treated unit, we want to estimate its untreated counterfactual:
- We'd like an estimate of $Y_i(0)$ for units with $D_i = 1$
- We can try to go for $Y(0; x)$ for a given $X_i = x$

Going beyond the exact matching estimator

What can we do when we have continuous X ?

- For each treated unit, we want to estimate its untreated counterfactual:
- We'd like an estimate of $Y_i(0)$ for units with $D_i = 1$
- We can try to go for $Y(0; x)$ for a given $X_i = x$

 What if we don't have any untreated people with $X_i = x$?

→ Find untreated units with X_i *close to* $X_i = x$

- With this population, we can simply take $\bar{Y}(0; x^{\text{close}})$
 - This is still flexible and non-parametric!

Going beyond the exact matching estimator

What can we do when we have continuous X ?

- For each treated unit, we want to estimate its untreated counterfactual:
- We'd like an estimate of $Y_i(0)$ for units with $D_i = 1$
- We can try to go for $Y(0; x)$ for a given $X_i = x$
- ☠ What if we don't have any untreated people with $X_i = x$?
- Find untreated units with X_i *close to* $X_i = x$
- With this population, we can simply take $\bar{Y}(0; x^{\text{close}})$
 - This is still flexible and non-parametric!

How do we define “close”?

Additional matching estimators

In datasets with continuous X s, we can:

- ① Match to “nearest neighbors”
 - ② Match within a bandwidth
- Different ways of getting “closeness”
- Non-parametric: no real functional form assumption on $Y(X)$

Nearest-neighbor matching

For each treated unit $i \in T$, we find its “nearest neighbor” in X :

- Take the untreated unit $j \in U$ with the smallest $|X_j - X_i|$
- Now your “counterfactual” is $\hat{Y}_i(0) = Y_j(0)$
- Repeat this for all treated units $i \in T$

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

Nearest-neighbor matching

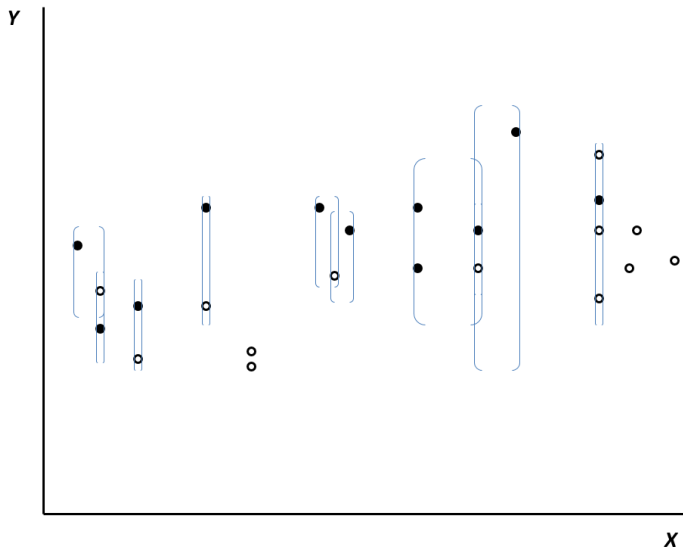
For each treated unit $i \in T$, we find its “nearest neighbor” in X :

- Take the untreated unit $j \in U$ with the smallest $|X_j - X_i|$
- Now your “counterfactual” is $\hat{Y}_i(0) = Y_j(0)$
- Repeat this for all treated units $i \in T$

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

- You can easily do this for an arbitrarily large K nearest neighbors
- With multiple neighbors, just average over the $Y_j(0)$'s to get $\hat{Y}_i(0)$
- Still not picking a functional form, but we are picking K

Getting “close” with nearest neighbors



Bandwidth matching

For each $i \in T$, we find $j \in U$ within a bandwidth h :

- Take all untreated units $j \in U$ with $x_j \in [X_i - h, X_i + h]$
- Now your “counterfactual” is $\hat{Y}_i(0) = \bar{Y}_j(0; X_i - h \leq X_j \leq X_i + h)$
- Repeat this for all treated units $i \in T$

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

Bandwidth matching

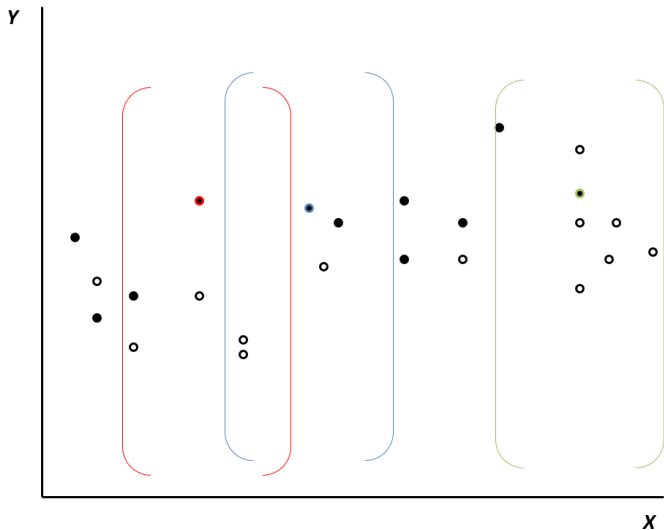
For each $i \in T$, we find $j \in U$ within a bandwidth h :

- Take all untreated units $j \in U$ with $x_j \in [X_i - h, X_i + h]$
- Now your “counterfactual” is $\hat{Y}_i(0) = \bar{Y}_j(0; X_i - h \leq X_j \leq X_i + h)$
- Repeat this for all treated units $i \in T$

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

- How do you choose a bandwidth?
 - Narrow: we'll get an accurate, but noisy estimate (similar X s, few observations)
 - Wide: we'll get an inaccurate, but precise estimate (different X s, many observations)
- We face a **bias-variance tradeoff**
- There are fancy tools for this (outside this class)

Getting “close” with bandwidths



A note on what we're estimating

For all three matching estimators, we can estimate ATE, ATT, or ATN:

- The trick is to make sure we know which one we're getting

- *Exact matching:*

- ATE: weight relative to the full sample: $\hat{\Delta}^{ATE} = \sum_{j=1}^{\# \text{ of cells}} \frac{N_j}{N} \hat{\Delta}_j$

- ATT: weight relative to the treated sample:

$$\hat{\Delta}^{ATT} = \sum_{k=1}^{\# \text{ of treated cells}} \frac{N_{k,T}}{N_T} \hat{\Delta}_k$$

- ATN: weight relative to the untreated sample:

$$\hat{\Delta}^{ATN} = \sum_{l=1}^{\# \text{ of untreated cells}} \frac{N_{l,U}}{N_U} \hat{\Delta}_l$$

- *Nearest neighbor and bandwidth matching:*

- ATT: For each treated unit, find untreated matches:

$$\hat{\tau}^{ATT} = \frac{1}{N_T} \sum_{i \in T} (Y_i(1) - \hat{Y}_i(0))$$

- ATN: For each untreated unit, find treated matches:

$$\hat{\tau}^{ATN} = \frac{1}{N_U} \sum_{i \in U} (\hat{Y}_i(1) - Y_i(0))$$

- ATE: Weight the ATT and ATN: $\hat{\tau}^{ATE} = \frac{N_T}{N_T + N_U} \hat{\tau}^{ATT} + \frac{N_U}{N_T + N_U} \hat{\tau}^{ATN}$

An example: Teacher value added

Policy issue:

- Improving student achievement is critical...but how?
- Teachers are probably a key component of this
- We want to measure teacher value-added

“Program” (more like an approach):

- Do good teachers improve student outcomes?
 - **Non-experimental** “program” (happening):
 - Teachers are (non-randomly) paired with students
- We don't have randomization, so we need an SOO design
- Compare similar students with different teachers

Estimating treatment effects of teachers

What happens to students with a better teacher (simplified)?

$$Y_i = \alpha + \tau \widehat{VA}_{ij} + \beta \mathbf{X}_i + \varepsilon_i$$

where

Y_i is a long-term outcome for student i

\widehat{VA}_{ij} is the value-added for teacher j who taught student i

\mathbf{X}_i are controls for student characteristics

ε_i is an error term

Estimating treatment effects of teachers

What happens to students with a better teacher (simplified)?

$$Y_i = \alpha + \tau \widehat{VA}_{ij} + \beta \mathbf{X}_i + \varepsilon_i$$

where

Y_i is a long-term outcome for student i

\widehat{VA}_{ij} is the value-added for teacher j who taught student i

\mathbf{X}_i are controls for student characteristics

ε_i is an error term

This is **not an experiment!** Two steps:

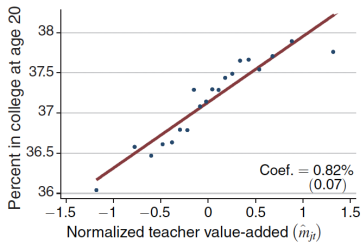
- 1 \widehat{VA}_{ij} comes from comparing teachers to themselves over time
 - Done using older data
 - 2 Estimate effects of VA on student outcomes
 - Students who were similar, but had differentially effective teachers
- **Identifying assumption:** After controlling for \mathbf{X}_i , students with good vs. bad teachers would have done similarly well

What's in \mathbf{X}_i ?

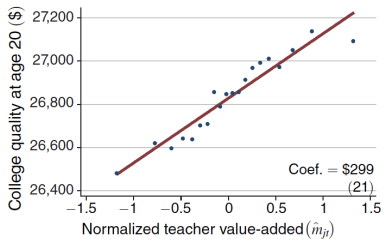
Control Vectors.—We construct residuals Y_{it} using separate models for each of the four subject-by-school-level cells. Within each of these groups, we regress raw outcomes Y_i^* on a vector of covariates \mathbf{X}_{it} with teacher fixed effects, as in (3), and compute residuals Y_{it} . We partition the control vector \mathbf{X}_{it} which we used to construct our baseline VA estimates into two components: student-level controls \mathbf{X}_{it}^I that vary across students within a class; and classroom-level controls \mathbf{X}_{ct} that vary only at the classroom level. The student-level control vector \mathbf{X}_{it}^I includes cubic polynomials in prior-year math and English scores, interacted with the student's grade level to permit flexibility in the persistence of test scores as students age. We also control for the following student level characteristics: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, free or reduced-price lunch, special education, and limited English. The class-level controls \mathbf{X}_{ct} consist of the following elements: (i) class size and class-type indicators (honors, remedial); (ii) cubics in class and school-grade means of prior-year test scores in math and English (defined based on those with non-missing prior scores) each interacted with grade; (iii) class and school-year means of all the individual covariates \mathbf{X}_{it}^I ; and (iv) grade and year dummies.

Teachers: Impacts on college

Panel A. College attendance at age 20



Panel B. College quality at age 20



Teachers: Different controls

TABLE 2— IMPACTS OF TEACHER VALUE-ADDED ON COLLEGE ATTENDANCE

	College at age 20 (%) (1)	College at age 20 (%) (2)	College at age 20 (%) (3)	College quality at age 20 (\$) (4)	College quality at age 20 (\$) (5)	College quality at age 20 (\$) (6)	High quality college (%) (7)	Four or more years of college, ages 18–22 (%) (8)
Teacher VA	0.82 (0.07)	0.71 (0.06)	0.74 (0.09)	298.63 (20.74)	265.82 (18.31)	266.17 (26.03)	0.72 (0.05)	0.79 (0.08)
Mean of dep. var.	37.22	37.22	37.09	26,837	26,837	26,798	13.41	24.59
Baseline controls	X	X	X	X	X	X	X	X
Parent chars. controls		X			X			
Lagged score controls			X			X		
Observations	4,170,905	4,170,905	3,130,855	4,167,571	4,167,571	3,128,478	4,167,571	3,030,878

Teachers: Different controls (earnings)

TABLE 3—IMPACTS OF TEACHER VALUE-ADDED ON EARNINGS

	Earnings at age 28 (\$) (1)	Earnings at age 28 (\$) (2)	Earnings at age 28 (\$) (3)	Working at age 28 (%) (4)	Total income at age 28 (\$) (5)	Wage growth ages 22–28 (\$) (6)
Teacher VA	349.84 (91.92)	285.55 (87.64)	308.98 (110.17)	0.38 (0.16)	353.83 (88.62)	286.20 (81.86)
Mean of dep. var.	21,256	21,256	21,468	68.09	22,108	11,454
Baseline controls	X	X	X	X	X	X
Parent chars. controls		X				
Lagged score controls			X			
Observations	650,965	650,965	510,309	650,965	650,965	650,943

Teachers: Quasi-experiment

TABLE 5—IMPACTS OF TEACHER VALUE-ADDED ON COLLEGE OUTCOMES: QUASI-EXPERIMENTAL ESTIMATES

	College attendance (%)				Predicted college attendance (%)
	(1)	(2)	(3)	(4)	(5)
<i>Panel A. College attendance at age 20</i>					
Teacher VA	0.86 (0.23)	0.73 (0.25)	0.67 (0.26)	1.20 (0.58)	0.02 (0.06)
Year FE	X				
School \times year FE		X	X	X	X
Lagged score controls			X		
Lead and lag changes in teacher VA			X		
Number of school \times grade \times subject \times year cells	33,167	33,167	26,857	8,711	33,167
Sample:	Full sample	Full sample	Full sample	No imputed scores	Full sample

Wrapping up SOO

We've covered the two main ways of doing SOO

- 1 Regression adjustment
 - Controlling for stuff
 - Makes parametric assumptions
- 2 Matching
 - Pairing observations
 - Less parametric

Wrapping up SOO

We've covered the two main ways of doing SOO

① Regression adjustment

- Controlling for stuff
- Makes parametric assumptions

② Matching

- Pairing observations
- Less parametric

A few last words:

- There are other, fancier ways to do this
- All make the extremely strong conditional independence assumption
- This is generally not reasonable in real life!
- We will end our treatment of SOO here

TL;DR:

- ① Selection on observables designs are **dubious**
- ② They require extremely strong assumptions!
- ③ But as a last resort, matching can be useful