

# Causal Inference I

MIXTAPE SESSION

---



# Roadmap

## Matching and weighting

- What is matching?

- Simple weighting

- Nearest neighbor matching

- Propensity scores

- Inverse probability weighting

  - Double robust

- Propensity score matching

- Coarsened exact matching

## Parameters of interest

- RCTs can recover the ATE
- DD and synth can identify the ATT
- Local average treatment effect can be identified using RDD and IV
- Can anything else recover the ATE and if so under what conditions?

## Defined causal parameters

Recall our definition of the ATE in terms of population mean differences in potential outcomes:

$$E[Y^1] - E[Y^0]$$

We work with samples, so we express the ATE sample analog:

$$\frac{1}{N} \sum_i [Y_i^1 - Y_i^0]$$

## Defined causal parameters

Our ATE estimator is a weighted average of the ATT and the ATU:

$$\hat{\delta}^{ATE} = \frac{N_T}{N} \hat{\delta}^{ATT} + \frac{N_C}{N} \hat{\delta}^{ATU}$$

where  $\hat{\delta}^{ATT} = \frac{1}{N_T} \sum_i \left[ Y_i^1 - \hat{Y}_i^0 \right]$ ;  $\hat{\delta}^{ATU} = \frac{1}{N_C} \sum_i \left[ \hat{Y}_i^1 - Y_i^0 \right]$ .

## Defined causal parameters

Our ATE estimator is a weighted average of the ATT and the ATU:

$$\hat{\delta}^{ATE} = \frac{N_T}{N} \hat{\delta}^{ATT} + \frac{N_C}{N} \hat{\delta}^{ATU}$$

where  $\hat{\delta}^{ATT} = \frac{1}{N_T} \sum_i \left[ Y_i^1 - \hat{Y}_i^0 \right]$ ;  $\hat{\delta}^{ATU} = \frac{1}{N_C} \sum_i \left[ \hat{Y}_i^1 - Y_i^0 \right]$ .

What is  $\hat{Y}_i^0$  for the treatment group? It's an **estimate** of the **missing** counterfactual. And this is its estimated value:

$$\hat{Y}_i^0 \equiv \frac{1}{Pr(D)} \sum_{j \in \{D_j=0\}} w_{ij} Y_j^0$$

- $w_{ij}$  is the weight unit  $j$  receives in predicting unit  $i$ 's counterfactual

# Estimating missing counterfactuals

$$\hat{Y}_i^0 \equiv \frac{1}{Pr(D)} \sum_{j \in \{D_j=0\}} w_{ij} Y_j^0$$

We will be estimating counterfactuals as **weighted** averages over outcomes from the other group (e.g., control) by incorporating observable covariates

- Matching and weighting methods will correct for relevant differences in observables, not just to do it, but because it closes the backdoor paths necessary for conditional independence to hold
- Estimators differ in how the weights,  $w_{ij}$ , are calculated
- Large part of the goal becomes **balance** on the conditioning set of covariates between treatment and control group needed to satisfy the backdoor criterion

## Big picture

- We are going to use weighting and matching methods to adjust for observable differences between treatment and control groups
- Personal philosophy: this **must be driven** by model like reasoning as this only has any hope if the backdoor criterion holds
- Personal observation: given the lack of comfort people have with modeling in general, it's not a good sign when you see them then use matching methods
- Let's define the terms first: what is a covariate?

# Observable Covariate

## Definition: Predetermined Covariates

Variable  $X$  is a predetermined covariate if for each individual  $i$ , the value of  $X_i$  does not depend on treatment status.

- Does not imply  $X$  and treatment status are independent
- Sometimes covariates do not change over time, but doesn't have to be
- Colliders are not covariates; anything the treatment can change is not a covariate ("bad controls", Angrist and Pischke MHE)

## Adjustment for Observables

- Simple weighting methods (e.g., subclassification)
- Exact matching methods (e.g., nearest neighbors)
- Approximate matching methods (e.g., propensity scores)
- Hybrid matching methods (e.g., coarsened exact matching)

# History of non-experimental matching

- A set of techniques for estimating ATE emerged in the 20th century from statistics and epidemiology, largely driven by public health concerns about smoking's connection to lung cancer
- "Weighting" methods to correct for covariate imbalance were developed to estimate ATE from non-experimental data
- In time these evolved into matching, and while conceptually they seem distinct, they are very similar
- Huge area, and it never really took off in economics despite some initial promise (Dehejia and Wahba 1999; 2002), but if you take Causal Inference Part 2, you'll see that the principles are there in difference-in-differences and synthetic control

# Smoking thought experiment

## **Does Smoking Cause Cancer?**

- Split a large enough population to gain enough power to detect causal effects into treatment and control
- Treatment spends their lives smoking a pack of day; control abstains
- Compare lung cancer rates between the two groups
- Low realism: can't really expect people to comply with a longterm experiment like this
- But important, so how did scientists proceed? Through weighting-based methods

Figure 1  
Lung Cancer at Autopsy: Combined Results from 18 Studies

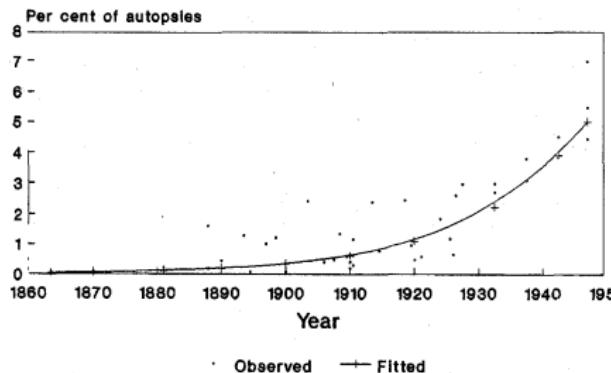


Figure 2(a)  
Mortality from Cancer of the Lung in Males

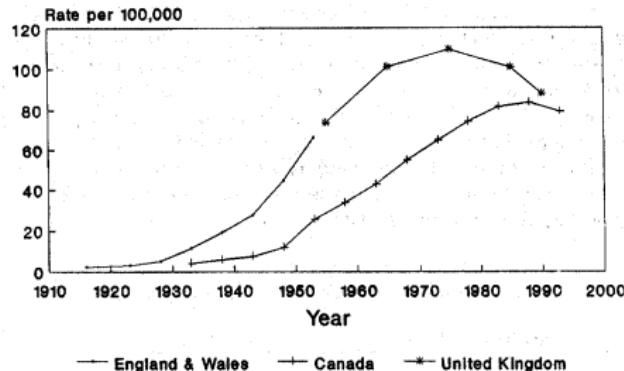


Figure 4  
Smoking and Lung Cancer Case-control Studies

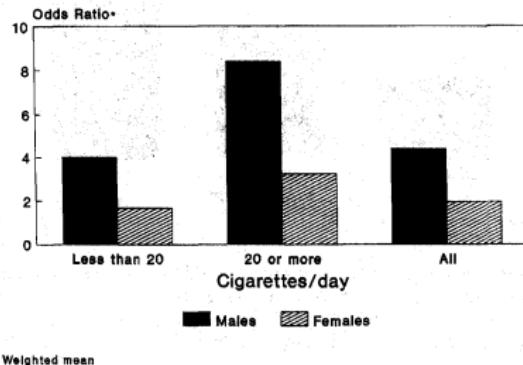
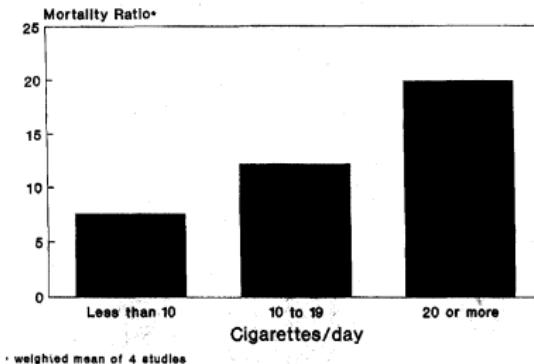
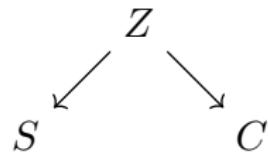


Figure 5  
Smoking and Lung cancer Cohort Studies in Males



# Does Smoking Cause Cancer?

Smoking,  $S$ , causes lung cancer,  $C$  ( $S \rightarrow C$ ) versus spurious correlation due to the  $Z$  confounder:



Legitimate criticism at the time, but incorrect in hindsight – hindsight is 20/20

## Nature of the criticism

Other criticisms came from giants like Joseph Berkson, Jerzy Neyman and Ronald Fisher

1. Correlation b/w smoking and lung cancer was spurious due to biased selection of subjects
2. Complaints about functional forms using “risk ratios” and “odds ratios”
3. Implausible magnitudes
4. Killer critique: *no experimental evidence* to incriminate smoking as a cause of lung cancer

# Fisher's confounding theory

- Fisher was a chain smoking pipe smoker, he died of cancer, and he was a paid expert witness for the tobacco industry
- He was also equally famous as a geneticist and arguing from logic, statistics and genetic evidence proposed a hypothetical confounding genome,  $Z$ , which introduced selection bias into contrasts of smokers and non-smokers ("perfect doctor")
- There was support for this: studies showed that cigarette smokers and non-smokers were different on observables – more extraverted than non-smokers and pipe smokers, differed in age, differed in income, differed in education, etc.

# Broken clocks are sometimes right

- Always easy to criticize someone when we look back with more information
- Evidence for the causal link was shallow:  
*"the [epidemiologists] turned out to be right, but only because bad logic does not necessarily lead to wrong conclusions." Robert Hooke (1983)*
- Scientists shifted to fixing their broken clocks for good

# Observable selection bias

Table: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Cigars in these data had much higher mortality rates in all three countries than non-smokers, but even cigarette smokers. Cigarette smokers in Canada have same mortality rates as non-smokers. Strange associations to us today, so imagine back then when the smoking-cancer hypothesis was not settled

# Non-smokers and smokers differ in age

*Table:* Mean ages, years (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

- Older people die at a higher rate, but for reasons other than just smoking cigars
- Could cigar smokers have higher observed death rates because they're older on average?
- How can we check this?

Force groups to have the same age distribution

Covariates are *not balanced* – their mean values differ for treatment and control group.

### **Option 1: Strata**

- Compare mortality rates across the different smoking groups but *within* each age group
- Compare within covariate strata and then combine differences to neutralize observed confounders

### **Option 2: Weights**

- Weight the data so that covariates are balanced, then compare mortality across treatment and control
- Which weights?

## Weighting the data

Divide the smoking group samples into age groups:

1. Calculate mortality rates separately for each age group *by treatment and control separately*
2. Construct “probability weights”: the proportion of each smoking group sample within a given age group
3. For treatment and control group, compute the weighted averages of the age groups mortality rates using the probability weights

This interestingly will balance the observed covariate, age, between treatment and control

## Simple weighting example

	Death rates		Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers	
Age 20-50	15	11	29	
Age 50-70	35	13	9	
Age +70	50	16	2	
Total		40	40	

Question: What is the average death rate for pipe smokers?

## Simple weighting example

	Death rates		Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers	
Age 20-50	15	11	29	
Age 50-70	35	13	9	
Age +70	50	16	2	
Total		40	40	

Question: What is the average death rate for pipe smokers?

$$15 \cdot \left(\frac{11}{40}\right) + 35 \cdot \left(\frac{13}{40}\right) + 50 \cdot \left(\frac{16}{40}\right) = 35.5$$

## Simple weighting example

	Death rates		Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers	
Age 20-50	15	11	29	
Age 50-70	35	13	9	
Age +70	50	16	2	
Total		40	40	

Counterfactual question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

## Simple weighting example

	Death rates		Number of	
	Pipe-smokers	Pipe-smokers	Non-smokers	
Age 20-50	15	11	29	
Age 50-70	35	13	9	
Age +70	50	16	2	
Total		40	40	

Counterfactual question: What would the average mortality rate be for pipe smokers if they had the same age distribution as the non-smokers?

$$15 \cdot \left(\frac{29}{40}\right) + 35 \cdot \left(\frac{9}{40}\right) + 50 \cdot \left(\frac{2}{40}\right) = 21.2$$

*Table:* Adjusted death rates using 3 age groups (Cochran 1968)

Smoking group	Canada	U.K.	U.S.
Non-smokers	20.2	11.3	13.5
Cigarettes	28.3	12.8	17.7
Cigars/pipes	21.2	12.0	14.2

# Assumptions, data and statistics

We need three things to estimate a causal effect

- **Assumptions:** what must we assume is true so that our models work with data?
- **Data:** what data with what covariates and outcomes do we need for this project?
- **Statistical models:** sometimes called “estimators” which are the calculators turning data into unbiased estimates of treatment effects (or not!)

## Assumption I: Independence

- Randomized treatment assignment guarantees “independence”

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

- Independence allows to estimate accurate causal effects through simple methods like differences in averages

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= \underbrace{E[Y^1|D=1] - E[Y^0|D=0]}_{\text{by the switching equation}} \\ &= \underbrace{E[Y^1] - E[Y^0]}_{\text{by independence}} \\ &= \underbrace{E[Y^1 - Y^0]}_{\text{ATE}} \end{aligned}$$

## Covariate distribution

- Just like independence implies balance on potential outcomes, it also implies balance on covariates which is called “common support”
- We saw this in our Thornton regressions: cash vouchers were not associated with being male, one’s age, etc.
- If we have balance on potential outcomes, that’s all we need but balance on covariates is often used to provide some evidence the randomization was done well

## Violations of independence

- Problem with smoking and cancer was smoking wasn't *randomly assigned* – it was chosen by the “perfect doctor” (i.e., self selection into smoking based on factors related to potential outcomes)
- When a treatment is “dependent” on potential outcomes, it means people smoke because they expect something is better when they smoke ( $Y^1$ ) than when they don't ( $Y^0$ ) introducing selection bias and potentially heterogenous treatment effect bias
- Naive comparisons can be deeply misleading – covariate adjustment can resolve this if “conditional independence” happens in the data

## Identifying assumption I: Conditional independence

$(Y_i^0, Y_i^1) \perp\!\!\!\perp D | X_i$ . There exists a set  $X$  of observable covariates such that after controlling for these covariates, treatment assignment is *independent of potential outcomes*.

- Conditional on  $X$ , treatment assignment is ‘as good as random’.
- ‘As good as random’ is English for “independent of potential outcomes” potential outcomes jargon

## Identifying assumption II: Common support

For ranges of  $X$ , there is a positive probability of being both treated and untreated

- Assumption requires that there are units in both treatment and control for the range of  $X$
- Common support ensures we can find similar enough donors in the control pool
- We can't check for balance on potential outcomes, because of switching equation eliminating counterfactuals, but we can check for common support

# Big picture of conditional independence

- You use these methods if you are confident in your DAG, **and** there exists a conditioning strategy that satisfies the backdoor criterion
- If you literally refuse to commit to a model, theory of DGP or DAG, then you cannot justify matching
- Matching is for experts – not because it's hard, but because experts are often the ones who *are willing* to write down a DAG; it rewards domain specific human capital
- But remember the audience and the progression away from models that Card (2014) mentioned ...

# Independence breaks down under perfect doctor reasoning

- Independence was violated if the treatment was assigned because we expected things to improve or not (“perfect doctor” reasoning)
- If you take an action because you think it helps and others don’t take the action because they don’t or can’t, then it is a violation of independence probably
- Keep telling yourself this phrase: “Firms don’t flip coins to set prices”. They set prices based on profit maximization or cost minimization or equivalent goals
- Selection bias then will be baked into those non-experimental observations and thus can’t be relied on without causal inference adjustments

# Conditional independence

- DAG reasoning can be a little unhelpful at times once we try to articulate the conditional independence assumption, so hear me out:
- Firms don't flip coins to set prices, but is there *some* random price setting conditional on observable factors (which only employees, managers and executives could possibly know about)?
- If they flip coins conditional on something, then that's "conditional independence". See – it's a strong belief.
- Conditional independence means that once we adjust for covariates, all remaining variation in treatment assignment had nothing to do with profit maximization or potential outcomes more generally but rather was as good as random

# Identification under conditional independence

Identification assumptions:

1.  $(Y^1, Y^0) \perp\!\!\!\perp D|X$  (conditional independence)
2.  $0 < Pr(D = 1|X) < 1$  with probability one (common support)

Comparing two individuals *who have the same values of  $X$* , treatment is independent of potential outcomes.

The second term implies we have people in treatment and control for every strata of  $X$

# Implications of assumptions

- Assumption 1 lets you plug  $Y$  for  $Y^j$  with the switching equation

$$\begin{aligned} E[Y^1 - Y^0 | X] &= E[Y^1 - Y^0 | X, D = 1] \\ &= E[Y | X, D = 1] - E[Y | X, D = 0] \end{aligned}$$

- Assumption 2 lets you properly weight over the covariate distribution

$$\begin{aligned} \delta_{ATE} &= E[Y^1 - Y^0] = E\left[E[Y^1 - Y^0 | X]\right] \\ &= \int E[Y^1 - Y^0 | X, D = 1] dPr(X) \\ &= \int (E[Y | X, D = 1] - E[Y | X, D = 0]) dPr(X) \end{aligned}$$

# Can we defend any conditional independence?

Other versions of conditional independence (and this shows up in diff-in-diff too)

1.  $Y^0 \perp\!\!\!\perp D|X$
2.  $Pr(D = 1|X) < 1$  (with  $Pr(D = 1) > 0$ )

Notice how there is only one potential outcome in the independence equation. That's okay. We can still then estimate the ATT (just not the ATE).

## ATT

Conditional independence of  $D$  with respect to  $Y^0$  conditional on  $X$ , but not  $Y^1$ , lets us recover the ATT with weights and the realized data

$$\begin{aligned}\delta_{ATT} &= E\left[E[Y^1 - Y^0 | D = 1, X]\right] \\ &= \int (E[Y|X, D = 1] - E[Y|X, D = 0]) dPr(X|D = 1)\end{aligned}$$

# Summarizing

Weighted averages under either assumption:

$$\delta_{ATE} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X)$$

$$\delta_{ATT} = \int (E[Y|X, D=1] - E[Y|X, D=0]) dPr(X|D=1)$$

ATE needs independence with respect to both potential outcomes; ATT only needs it with respect to  $Y^0$ .

## Weighting by Age ( $K = 2$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is  $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N^k}{N} \right)$ ?

## Weighting by Age ( $K = 2$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is  $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left(\frac{N^k}{N}\right)$ ?

$$4 \cdot \left(\frac{13}{30}\right) + 6 \cdot \left(\frac{17}{30}\right) = 5.13$$

## Weighting by Age ( $K = 2$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is  $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N_T^k}{N_T} \right)$ ?

## Weighting by Age ( $K = 2$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old	28	24	4	3	10
Young	22	16	6	7	10
Total				10	20

Question: What is  $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N_T^k}{N_T} \right)$ ?

$$4 \cdot \left( \frac{3}{10} \right) + 6 \cdot \left( \frac{7}{10} \right) = 5.4$$

## Weighting by Age and Gender ( $K = 4$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is  $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N^k}{N} \right)$ ?

# Weighting by Age and Gender ( $K = 4$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Problem: What is  $\widehat{\delta_{ATE}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N^k}{N} \right)$ ?

Not identified! What went wrong?

## Weighting by Age and Gender ( $K = 4$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is  $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N_T^k}{N_T} \right)$ ?

## Weighting by Age and Gender ( $K = 4$ )

$X_k$	Death Rate			Number of	
	Smokers	Non-smokers	Diff.	Smokers	Non-smokers
Old Males	28	22	4	3	7
Old Females		24			3
Young Males	21	16	5	3	4
Young Females	23	17	6	4	6
Total				10	20

Question: What is  $\widehat{\delta_{ATT}} = \sum_{k=1}^K (\bar{Y}^{1,k} - \bar{Y}^{0,k}) \cdot \left( \frac{N_T^k}{N_T} \right)$ ?

$$4 \cdot \left( \frac{3}{10} \right) + 5 \cdot \left( \frac{3}{10} \right) + 6 \cdot \left( \frac{4}{10} \right) = 5.1$$

# Curse of Dimensionality

- Simple weighting methods may become less feasible in finite samples as the number of covariates grows (e.g.,  $K = 4$  was too many for this sample)
- Assume we have  $k$  covariates and we divide each into 3 coarse categories (e.g., age: young, middle age, old; income: low, medium, high, etc.)
- The number of sub classification cells (or “strata”) is  $3^k$ . For  $k = 10$ , then it’s  $3^{10} = 59,049$

# Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both. If so, we cannot use this kind of simple weighting (e.g., subclassification, stratification).
- Simple weighting methods is also a problem if the cells are “too coarse”.
- We can always use “finer” classifications, but finer cells worsens the dimensional problem, so we don’t gain much from that. ex: using 10 variables and 5 categories for each, we get  $5^{10} = 9,765,625$ .
- Matching methods really force us to see these curses; they’re often hidden from OLS because OLS doesn’t tell us it is just doing various extrapolations

## To Look Like Someone Else

- Not sure of the exact history of matching, but the idea is super intuitive
- What if we could make synthetic xerox copies of ourselves in counterfactual states?
- Matching is to basically find units in the control group, or treatment group, that “look like me” **on observables**
- But remember – it’s driven by DAG reasoning; you only need to look minimally “like” the other person to satisfy backdoor

# Nearest Neighbor Matching

- See Abadie and Imbens (2006). “Large sample properties of matching estimators for average treatment effects”. *Econometrica*
- We could also estimate  $\delta_{ATT}$  by *imputing* the missing potential outcome of each treatment unit  $i$  using the observed outcome from that outcome’s “nearest” neighbor  $j$  in the control set

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$$

where  $Y_{j(i)}$  is the observed outcome of a control unit such that  $X_{j(i)}$  is the **closest** value to  $X_i$  among all of the control observations (eg match on  $X$ )

# Matching

- We could also use the average observed outcome over  $M$  closest matches:

$$\delta_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \left[ \frac{1}{M} \sum_{m=1}^M Y_{j_m(1)} \right] \right)$$

- Works well when we can find good matches for each treatment group unit, so  $M$  is usually defined to be small (i.e.,  $M = 1$  or  $M = 2$ )

# Matching

- We can also use matching to estimate  $\delta_{ATE}$ . In that case, we match in both directions:
  1. If observation  $i$  is treated, we impute  $Y_i^0$  using the control matches,  $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
  2. If observation  $i$  is control, we impute  $Y_i^1$  using the treatment matches,  $\{Y_{j_1(i)}, \dots, Y_{j_M(i)}\}$
- The estimator is:

$$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^N (2D_i - 1) \left[ Y_i - \left( \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} \right) \right]$$

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_i$	$X_i$
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_i$	$X_i$
1	6	?	1	3
2	1	?	1	1
3	0	?	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\widehat{\delta_{ATT}} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

Match and plug in!

## Matching example with single covariate

$i$	$Y_i^1$	$Y_i^0$	$D_I$	$X_i$
1	6	9	1	3
2	1	0	1	1
3	0	9	1	10
4		0	0	2
5		9	0	3
6		1	0	-2
7		1	0	-4

Question: What is  $\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)})$ ?

$$\widehat{\delta}_{ATT} = \frac{1}{3} \cdot (6 - 9) + \frac{1}{3} \cdot (1 - 0) + \frac{1}{3} \cdot (0 - 9) = -3.7$$

## A Training Example

Trainees			Non-Trainees		
unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900
2	34	10200	2	50	31000
3	29	14400	3	30	21000
4	25	20800	4	27	9300
5	29	6100	5	54	41100
6	23	28600	6	48	29800
7	33	21900	7	39	42000
8	27	28800	8	28	8800
9	31	20300	9	24	25500
10	26	28100	10	33	15500
11	25	9400	11	26	400
12	27	14300	12	31	26600
13	29	12500	13	26	16500
14	24	19700	14	34	24200
15	25	10100	15	25	23300
16	43	10700	16	24	9700
17	28	11500	17	29	6200
18	27	10700	18	35	30200
19	28	16300	19	32	17800
Average:		28.5	16426	20	23
			21	32	25900
			Average:		20724

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900			
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000			
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000			
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300			
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100			
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800			
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000			
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800			
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500			
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500			
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400			
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:		20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600			
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500			
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200			
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300			
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700			
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:		20724	Average:		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200			
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:	28.5	16426	20	23	9500			
			21	32	25900			
			Average:	33	20724	Average:		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200			
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800			
Average:		28.5	16426	20	23	9500	Average:	
			21	32	25900			
			Average:		33	20724		

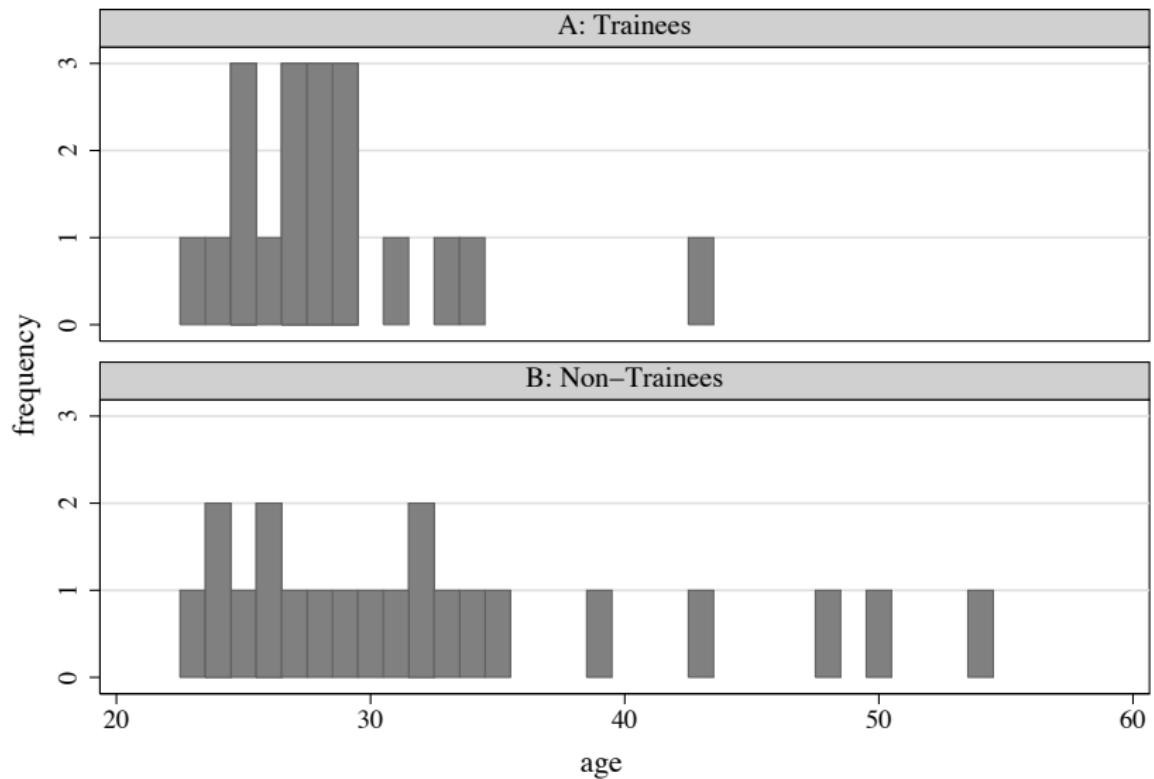
# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:		
			21	32	25900			
			Average:		20724			

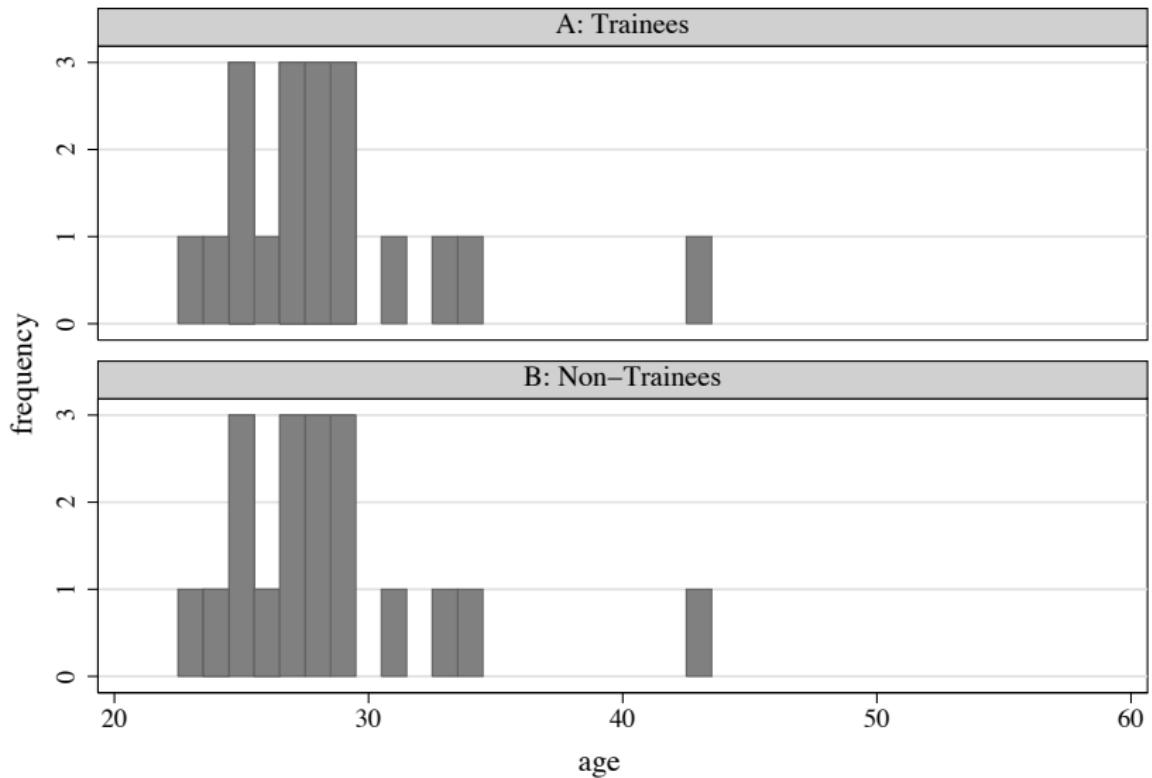
# A Training Example

Trainees			Non-Trainees			Matched Sample		
unit	age	earnings	unit	age	earnings	unit	age	earnings
1	28	17700	1	43	20900	8	28	8800
2	34	10200	2	50	31000	14	34	24200
3	29	14400	3	30	21000	17	29	6200
4	25	20800	4	27	9300	15	25	23300
5	29	6100	5	54	41100	17	29	6200
6	23	28600	6	48	29800	20	23	9500
7	33	21900	7	39	42000	10	33	15500
8	27	28800	8	28	8800	4	27	9300
9	31	20300	9	24	25500	12	31	26600
10	26	28100	10	33	15500	11,13	26	8450
11	25	9400	11	26	400	15	25	23300
12	27	14300	12	31	26600	4	27	9300
13	29	12500	13	26	16500	17	29	6200
14	24	19700	14	34	24200	9,16	24	17700
15	25	10100	15	25	23300	15	25	23300
16	43	10700	16	24	9700	1	43	20900
17	28	11500	17	29	6200	8	28	8800
18	27	10700	18	35	30200	4	27	9300
19	28	16300	19	32	17800	8	28	8800
Average:	28.5	16426	20	23	9500	Average:	28.5	13982
			21	32	25900			
			Average:		20724			

# Age Distribution: Before Matching



# Age Distribution: After Matching



## Training Effect Estimates

Difference in average earnings between trainees and non-trainees

- Before matching

$$16426 - 20724 = -4298$$

- After matching:

$$16426 - 13982 = 2444$$

## Alternative distance metric: Euclidean distance

When the vector of matching covariates,  $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{pmatrix}$  has more than one dimension ( $k > 1$ ) we will need a new definition of **distance** to measure “closeness”.

## Alternative distance metric: Euclidean distance

### Definition: Euclidean distance

$$\begin{aligned} \|X_i - X_j\| &= \sqrt{(X_i - X_j)'(X_i - X_j)} \\ &= \sqrt{\sum_{n=1}^k (X_{ni} - X_{nj})^2} \end{aligned}$$

**Comment:** The Euclidean distance is not invariant to changes in the scale of the  $X$ 's. For this reason, alternative distance metrics that are invariant to changes in scale are used

# Normalized Euclidean distance

Definition: Normalized Euclidean distance

A commonly used distance is the normalized Euclidean distance:

$$||X_i - X_j|| = \sqrt{(X_i - X_j)' \hat{V}^{-1} (X_i - X_j)}$$

where

$$\hat{V}^{-1} = \text{diag}(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \dots, \hat{\sigma}_k^2)$$

## Normalized Euclidean distance

- Notice that the normalized Euclidean distance is equal to:

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \frac{(X_{ni} - X_{nj})}{\hat{\sigma}_n^2}}$$

- Thus, if there are changes in the scale of  $X_{ni}$ , these changes also affect  $\hat{\sigma}_n^2$ , and the normalized Euclidean distance does not change

# Mahalanobis distance

## Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

$$\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}_X^{-1} (X_i - X_j)}$$

where  $\hat{\Sigma}_X$  is the sample variance-covariance matrix of  $X$ .

## Arbitrary weights

Or, you could just create your own arbitrary weights

$$\|X_i - X_j\| = \sqrt{\sum_{n=1}^k \omega_n \cdot (X_{ni} - X_{nj})^2}$$

(with all  $\omega_n \geq 0$ ) so that we assign large  $\omega_n$ 's to those covariates that we want to match particularly well.

# Matching and the Curse of Dimensionality

Dimensionality creates headaches for us in matching.

- **Bad news:** Matching discrepancies  $\|X_i - X_{j(i)}\|$  tend to increase with  $k$ , the dimension of  $X$
- **Good news:** Matching discrepancies converge to zero ...
- **Bad news:** ... but they converge very slow if  $k$  is large
- **Good news:** Mathematically, it can be shown that  $\|X_i - X_{j(i)}\|$  converges to zero at the same rate as  $\frac{1}{N^{\frac{1}{k}}}$
- **Bad news:** It's hard to find good matches when  $X$  has a large dimension: you need many observations if  $k$  is big.

## Deriving the matching bias

$$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where each  $i$  and  $j(i)$  units are matched,  $X_i \approx X_{j(i)}$  and  $D_{j(i)} = 0$ .

Define potential outcomes and switching eq.

$$\mu^0(x) = E[Y|X = x, D = 0] = E[Y^0|X = x],$$

$$\mu^1(x) = E[Y|X = x, D = 1] = E[Y^1|X = x],$$

$$Y_i = \mu^{D_i}(X_i) + \varepsilon_i$$

# Deriving the matching bias

Substitute and distribute terms

$$\begin{aligned}\hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (Y_i - Y_{j(i)}) \\ &= \frac{1}{N_T} \sum_{D_i=1} [(\mu^1(X_i) + \varepsilon_i) - (\mu^0(X_{j(i)}) + \varepsilon_{j(i)})] \\ &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)})) + \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

# Deriving the matching bias

Difference between sample estimate and population parameter is:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_{j(i)}) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)})\end{aligned}$$

Algebraic manipulation and simplification:

$$\begin{aligned}\widehat{\delta}_{ATT} - \delta_{ATT} &= \frac{1}{N_T} \sum_{D_i=1} (\mu^1(X_i) - \mu^0(X_i) - \delta_{ATT}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &+ \frac{1}{N_T} \sum_{D_i=1} (\mu^0(X_i) - \mu^0(X_{j(i)})) .\end{aligned}$$

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ .

## Deriving the matching bias

Note  $\widehat{\delta}_{ATT} - \delta_{ATT} \rightarrow 0$  as  $N \rightarrow \infty$ . However,

$$E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right] = E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right].$$

Now consider the implications if  $k$  is large:

- The difference between  $X_i$  and  $X_{j(i)}$  converges to zero very slowly
- The difference  $\mu^0(X_i) - \mu^0(X_{j(i)})$  converges to zero very slowly
- $E\left[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1\right]$  may not converge to zero and can be very large!
- $E\left[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})\right]$  may not converge to zero because the bias of the matching discrepancy is dominating the matching estimator!

Bias is often an issue when we match in many dimensions

## Solutions to matching bias problem

The bias of the matching estimator is caused by large matching discrepancies  $\|X_i - X_{j(i)}\|$ . The curse of dimensionality virtually guarantees this. However:

1. But the matching discrepancies are observed. We can always check in the data how well we're matching the covariates.
2. For  $\widehat{\delta}_{ATT}$  we can sometimes make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making  $N_C$  large).
3. If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques

## Matching with bias correction

- Each treated observation contributes

$$\mu^0(X_i) - \mu^0(X_{j(i)})$$

to the bias.

- Bias-corrected (BC) matching:

$$\hat{\delta}_{ATT}^{BC} = \frac{1}{N_T} \sum_{D_i=1} \left[ (Y_i - Y_{j(i)}) - (\widehat{\mu^0}(X_i) - \widehat{\mu^0}(X_{j(i)})) \right]$$

where  $\widehat{\mu^0}(X)$  is an estimate of  $E[Y|X = x, D = 0]$ . For example using OLS.

- Under some conditions, the bias correction eliminates the bias of the matching estimator without affecting the variance.

## Bias adjustment in matched data

---

unit	Potential Outcome		$D_i$	$X_i$
	under Treatment	under Control		
$i$	$Y_i^1$	$Y_i^0$		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

## Bias adjustment in matched data

---

unit	Potential Outcome			$X_i$
	under Treatment	under Control	$D_i$	
$i$	$Y_i^1$	$Y_i^0$		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\hat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate  $\widehat{\mu}^0(X) = \hat{\beta}_0 + \hat{\beta}_1 X = 2 + X$

## Bias adjustment in matched data

unit <i>i</i>	Potential Outcome		$D_i$	$X_i$
	under Treatment	under Control		
1	10	8	1	3
2	4	1	1	1
3	10	9	1	10
4		8	0	4
5		1	0	0
6		9	0	8

$$\widehat{\delta}_{ATT} = \frac{10 - 8}{3} + \frac{4 - 1}{3} + \frac{10 - 9}{3} = 2$$

For the bias correction, estimate  $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\begin{aligned}\widehat{\delta}_{ATT} &= \frac{(10 - 8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4 - 1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3} \\ &+ \frac{(10 - 9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33\end{aligned}$$

## Matching bias: Implications for practice

Bias arises because of the effect of large matching discrepancies on  $\mu^0(X_i) - \mu^0(X_{j(i)})$ . To minimize matching discrepancies:

1. Use a small  $M$  (e.g.,  $M = 1$ ). Larger values of  $M$  produce large matching discrepancies.
2. Use matching with replacement. Because matching with replacement can use untreated units as a match more than once, matching with replacement produces smaller matching discrepancies than matching without replacement.
3. Try to match covariates with a large effect on  $\mu^0(\cdot)$  particularly well.

# Large sample distribution for matching estimators

- Matching estimators have a Normal distribution in large samples (provided the bias is small):

$$\sqrt{N_T}(\hat{\delta}_{ATT} - \delta_{ATT}) \xrightarrow{d} N(0, \sigma_{ATT}^2)$$

- For matching without replacement, the “usual” variance estimator:

$$\hat{\sigma}_{ATT}^2 = \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \hat{\delta}_{ATT} \right)^2,$$

is valid.

# Large sample distribution for matching estimators

- For matching with replacement:

$$\begin{aligned}\widehat{\sigma}_{ATT}^2 &= \frac{1}{N_T} \sum_{D_i=1} \left( Y_i - \frac{1}{M} \sum_{m=1}^M Y_{j_m(i)} - \widehat{\delta}_{ATT} \right)^2 \\ &+ \frac{1}{N_T} \sum_{D_i=0} \left( \frac{K_i(K_i-1)}{M^2} \right) \widehat{var}(\varepsilon|X_i, D_i = 0)\end{aligned}$$

where  $K_i$  is the number of times observation  $i$  is used as a match.

- $\widehat{var}(Y_i|X_i, D_i = 0)$  can be estimated also by matching. For example, take two observations with  $D_i = D_j = 0$  and  $X_i \approx X_j$ , then

$$\widehat{var}(Y_i|X_i, D_i = 0) = \frac{(Y_i - Y_j)^2}{2}$$

is an unbiased estimator of  $\widehat{var}(\varepsilon_i|X_i, D_i = 0)$

## Avoiding dimensionality problems

- Curse of dimensionality makes matching on  $K$  covariates challenging
- Rubin (1977) and Rosenbaum and Rubin (1983) develop a method that can contain those  $K$  covariates used for adjusting
- Insofar as treatment is random conditional on  $K$  covariates, then one can use the propensity score to adjust for confounders

# Least squares

- OLS is best linear predictor and approximation to the conditional expectation function
- But if probability of treatment is nonlinear, this conditional mean may be less informative
- Propensity scores relax the linearity assumption and have other advantages

## The Idea behind propensity scores

- Earlier we matched on  $X$ 's to compare units “near” one another based on some distance but matching discrepancies and sparseness created problems
- Propensity scores summarize covariate information about treatment selection into a single number bounded between 0 and 1 (i.e., a probability)
- Now we compare units with similar *estimated probabilities* of treatment
- And once we adjust using the propensity score, we no longer need to adjust for  $X$

# Formal Definition

## Definition of Propensity score

A propensity score is a number bounded between 0 and 1 measuring the probability of treatment assignment conditional on a vector of confounding variables:  $p(X) = Pr(D = 1|X)$

Two Necessary Identification Assumptions:

1.  $(Y^0, Y^1) \perp\!\!\!\perp D|X$  (CIA)
2.  $0 < Pr(D = 1|X) < 1$  (common support)

## Steps

1. Estimate the propensity score using logit/probit
2. Estimate a particular ATE incorporating the propensity score using stratification, imputation, regression, or inverse probability weighting
3. Estimate standard errors

# Estimating the propensity score

- Estimate the conditional probability of treatment using probit or logit model

$$Pr(D_i = 1|X_i) = F(\beta X_i)$$

- Use the estimated coefficients to calculate the propensity score for each unit  $i$

$$\hat{\rho}_i = \hat{\beta} X_i$$

- Propensity score is the predicted conditional probability of treatment, or the fitted value for each unit – it is the likelihood you'd be in the treatment group given your covariates
- Frequentist probability in other words

# Identification

- A group of unit's average treatment effect may depend on some characteristic,  $X$

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

- CIA allow us to substitute

$$E[Y_i | D_i = 1, X_i = x] = E[Y_i^1 | D_i = 1, X_i = x]$$

and similar for other term  $Y^0$  using a switching equation

- Common support allows us to estimate both terms

## Propensity score theorem

If  $(Y^1, Y^0) \perp\!\!\!\perp D|X$  (CIA), then  $(Y^1, Y^0) \perp\!\!\!\perp D|\rho(X)$  where  $\rho(X) = Pr(D = 1|X)$ , the propensity score

- Conditioning on the propensity score is enough to have independence between  $D$  and  $(Y^1, Y^0)$  (Rosenbaum and Rubin 1983)
- Valuable theorem because of dimension reduction and convergence rate issues which can introduce biases
- **Big picture:** You can toss  $X$  out if you have  $\hat{\rho}$  because all information from  $X$  have been absorbed into  $\hat{\rho}$

## Proof

- Before diving into the proof, first recognize that

$$Pr(D = 1|Y^0, Y^1, \rho(X)) = E[D|Y^0, Y^1, \rho(X)]$$

because

$$\begin{aligned} E[D|Y^0, Y^1, \rho(X)] &= 1 \times Pr(D = 1|Y^0, Y^1, \rho(X)) \\ &\quad + 0 \times Pr(D = 0|Y^0, Y^1, \rho(X)) \end{aligned}$$

and the second term cancels out.

## Proof.

Assume  $(Y^1, Y^0) \perp\!\!\!\perp D|X$  (CIA). Then:

$$\begin{aligned} Pr(D = 1|Y^1, Y^0, \rho(X)) &= \underbrace{E[D|Y^1, Y^0, \rho(X)]}_{\text{See previous slide}} \\ &= \underbrace{E[E[D|Y^1, Y^0, \rho(X), X]|Y^1, Y^0, \rho(X)]}_{\text{by LIE}} \\ &= \underbrace{E[E[D|Y^1, Y^0, X]|Y^1, Y^0, \rho(X)]}_{\text{Given } X, \text{ we know } p(X)} \\ &= \underbrace{E[E[D|X]|Y^1, Y^0, \rho(X)]}_{\text{by CIA}} \\ &= \underbrace{E[\rho(X)|Y^1, Y^0, \rho(X)]}_{\text{propensity score definition}} \\ &= \rho(X) \end{aligned}$$



## Similar proof

We also can show that the probability of treatment conditional on the propensity score is the propensity score using a similar argument:

$$\begin{aligned} Pr(D = 1|\rho(X)) &= \underbrace{E[D|\rho(X)]}_{\text{Previous slide}} \\ &= \underbrace{E[E[D|X]|\rho(X)]}_{\text{LIE}} \\ &= \underbrace{E[p(X)|\rho(X)]}_{\text{definition}} \\ &= \rho(X) \end{aligned}$$

and  $Pr(D = 1|Y^1, Y^0, \rho(X)) = Pr(D = 1|\rho(X))$  by CIA

# Unbiased estimation of the ATE

Exact methods to do this to be discussed later, but until then, we can say this:

## Corollary: Estimating the ATE

If  $(Y^1, Y^0) \perp\!\!\!\perp D|X$ , we can estimate average treatment effects:

$$E[Y^1 - Y^0 | \rho(X)] = E[Y|D = 1, \rho(X)] - E[Y|D = 0, \rho(X)]$$

## Balancing property

- Because the propensity score is a function of  $X$ , we know:

$$\begin{aligned} Pr(D = 1|X, \rho(X)) &= Pr(D = 1|X) \\ &= \rho(X) \end{aligned}$$

- Conditional on  $\rho(X)$ , the probability that  $D = 1$  does not depend on  $X$ .
- $D$  and  $X$  are independent conditional on  $\rho(X)$ :

$$D \perp\!\!\!\perp X | \rho(X)$$

## Balancing property

- So we obtain the **balancing property** of the propensity score:

$$Pr(X|D = 1, p(X)) = Pr(X|D = 0, p(X))$$

conditional on the property score, the distribution of the covariates is the same for treatment and control group units

- We can use this to check if our estimated propensity score actually produces balance:

$$Pr(X|D = 1, \hat{p}(X)) = Pr(X|D = 0, \hat{p}(X))$$

## Propensity score theorem

- This theorem tells us the *only* covariate we need to adjust for is the conditional probability of treatment itself (i.e., the propensity score)
- It does not tell us which method we should use to do that adjustment, though, which is an estimation question
- There are options: inverse probability weighting, forms of imputation, stratification, and sometimes even regressions will incorporate the score as weights

## Checking the common support assumption

- We can summarize the propensity scores in the treatment and control group and count how many units are off-support
- Crump, et al. (2009) offer a rule of thumb: keep scores on interval [0.1,0.9].
- Tossing out observations beyond those min and max scores
- A histogram of propensity scores by treatment and control group also highlights the overlap problem; software also can help such as `teffects overlap`

# Inverse probability weighting

- I really like the simple method of inverse probability weighting aesthetically because there are no black boxes; it's all non-parametric averaging done through a particular kind of weights based on the propensity score
- IPW involves fewer implementation choices like number of neighbors, common support, etc.
- And because IPW is a smooth estimator, the bootstrap is valid for inference unlike covariate nearest neighbor matching which Abadie and Imbens (2008) show is not valid

# Inverse probability weighting

- IPW is basically a reweighting of the outcomes by the propensity score developed in Robins and Rotnitzky (1995), Imbens (2000), Hirano and Imbens (2001)
- The weights can be expressed in two ways – without normalization (Horvitz and Thompson 1952) or normalized (Hajek 1971) – the difference being how well either approach can handle extreme values of the propensity score; the differences come out of the survey sampling literature
- Notation is far far scarier than in fact what we are doing, so I'll show you this in a Stata and R simulation to help pin down the intuition a little better
- We'll start with the basic idea using the Horvitz and Thompson (1952) expression of the weights as it's not as messy.

# Inverse Probability Weighting

## Proposition

If  $Y^1, Y^0 \perp\!\!\!\perp D|X$ , then

$$\begin{aligned}\delta_{ATE} &= E[Y^1 - Y^0] \\ &= E\left[Y \cdot \frac{D - \rho(X)}{\rho(X) \cdot (1 - \rho(X))}\right] \\ \delta_{ATT} &= E[Y^1 - Y^0 | D = 1] \\ &= \frac{1}{Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - \rho(X)}{1 - \rho(X)}\right]\end{aligned}$$

## IPW Proof

Proof.

$$\begin{aligned} E \left[ Y \cdot \frac{D - \rho(X)}{\rho(X)(1 - \rho(X))} \middle| X \right] &= E \left[ \frac{Y}{\rho(X)} \middle| X, D = 1 \right] \rho(X) \\ &\quad + E \left[ \frac{-Y}{1 - \rho(X)} \middle| X, D = 0 \right] (1 - \rho(X)) \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

and the results follow from integrating over  $P(X)$  and  $P(X|D = 1)$ . □

# Weighting on the propensity score

Previous formulas used population concepts. Switching to samples, we use a two-step estimator:

1. Estimate the propensity score:  $\hat{\rho}(X)$
2. Use estimated score to produce analog estimators. Let  $\hat{\delta}_{ATE}$  and  $\hat{\delta}_{ATT}$  be an estimate of the  $ATE$  and  $ATT$  parameter:

$$\begin{aligned}\hat{\delta}_{ATE} &= \frac{1}{N} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{\rho}(X_i)}{\hat{\rho}(X_i) \cdot (1 - \hat{\rho}(X_i))} \\ \hat{\delta}_{ATT} &= \frac{1}{N_T} \sum_{i=1}^N Y_i \cdot \frac{D_i - \hat{\rho}(X_i)}{1 - \hat{\rho}(X_i)}\end{aligned}$$

## Weighting on the propensity score

Standard errors can be constructed a few different ways:

- We need to adjust the standard errors for first-step estimation of  $\rho(X)$
- Parameteric first step: Newey and McFadden (1994)
- Non-parametric first step: Newey (1994)
- Or bootstrap the entire two-step procedure (Adudumilli 2018 and Bodory et al. 2020)

## Implementation with software

- I like estimating with IPW manually because I like being reminded how simple a procedure it is
- But Stata's `-teffects-` and R's `-ipw-` do it too, and `-teffects-` uses the Hajek normalization weights which will produce identical estimates to my program
- My programs don't do the inference, but I think that would be fun and easy to do using the bootstrap
- Let's look at it real quickly now with an example from LaLonde's 1986 paper on the NSW job trainings program (which I'll discuss again soon)

# Double robust estimators

- Lots of papers: Robins and Rotnizky (1995) originally, Hirano and Imbens (2001), etc.
- Basic idea is you are going to control for covariates twice: through regression and then through the propensity score
- We say that estimators combining regression with IPW are double robust so long as
  - The regression for the outcome is properly specified, or
  - The propensity score is properly specified
- Hence the name “double robust”. We give ourselves two chances to get it right (either/or not both/and)

## Estimation of outcome model

$$y_i = \alpha_0 + X_i\beta + \tilde{\alpha}_1 D_i + \theta_0 \frac{D_i}{\widehat{\rho(X_i)}} + \theta_1 \frac{1 - D_i}{1 - \widehat{\rho(X_i)}} + \tilde{\varepsilon}_i$$

## **Propensity score matching**

- Matching, or what I like to call “imputation”, is another way that utilizes the  $\hat{p}$
- They all use the same first stage, but differ on their second and third stages
- Part of the second stage may be imposing common support through “trimming”, but for different reasons because now this idea of distance is entering and maybe you think some units are “too far away” to be relevant counterfactuals

## Standard matching strategy

- Pair each treatment unit  $i$  with one or more *comparable* control group unit  $j$ , where comparability is in terms of proximity to the estimated propensity score
- Impute the unit's missing counterfactual outcome  $Y_{i(j)}$  based on the unit or units chosen in the previous step
- If more than one are “nearest neighbors”, then use the neighbors' weighted outcomes

$$Y_{i(j)} = \sum_{j \in C(i)} w_{ij} Y_j$$

where  $C(i)$  is the set of neighbors with  $W = 0$  of the treatment unit  $i$  and  $w_{ij}$  is the weight of control group units  $j$  with  $\sum_{j \in C(i)} w_{ij} = 1$

# Imputing the counterfactuals

A parameter of interest:

$$E[Y_i^1|D_i = 1] - E[Y_i^0|D_i = 1]$$

We estimate it as follows

$$\widehat{ATT} = \frac{1}{N_T} = \sum_{i:W_i=1} \left[ Y_i - Y_{i(j)} \right]$$

where  $N_T$  is the number of matched treatment units in the sample.  
Note the difference between *imputation* and weighting

# Matching methods

- The probability of observing two units with exactly the same propensity score is in principle zero because  $p(x)$  is continuous
- Several matching methods have been proposed in the literature, but the most widely used are:
  - Stratification matching
  - Nearest-neighbor matching (with or without caliper)
  - Radius matching
  - Kernel matching
- Typically, one treatment unit  $i$  is matched to several control units  $j$ , but sometimes one-to-one matching is used

# Stratification

- Stratification is used to force covariate balance by finding strata where there is no difference in mean covariate values.
- You then use those strata to calculate within differences in means and sum over properly weighted strata. See Becker and Ichino (2002)
- Stratification is a brute force method for imposing balance by grouping the data and testing for differences in covariate means
- It's actually kind of similar to coarsened exact matching, only using the propensity score for the "stratification" not the covariates

# Stratification: Achieving Balance

The algorithm is brute force covariate balancing

1. Sort the data by propensity score and divide into groups of observations with similar propensity scores (e.g., percentiles)
2. Within each group, test (using a t-test) whether the means of the covariates ( $X$ ) are equal between treatment and control
3. If so, then stop. If not, it means the covariates aren't balanced *within that group*. Divide the group in half and repeat
4. If a particular covariate is unbalanced for multiple groups, modify the initial logit or probit equation by including higher order terms and/or interactions with that covariate and repeat

Historically this could be done with `-pscore2.ado-` or manually oneself if they felt so inclined, but it was dropped with `-teffects-`

# Nearest Neighbor

Pretty similar to covariate matching. Formula is

$$ATT^{NN} = \frac{1}{N_T} \sum_{i:W_i=1} \left[ Y_i - \sum_{j \in C(i)_M} w_{ij} Y_j \right]$$

- $N_T$  is the number of Treatment units  $i$
- $w_{ij}$  is equal to  $\frac{1}{N_C}$  if  $j$  is a control unit and zero otherwise;  $N_C$  is number of control units  $j$
- And unit  $j$  is chosen as a control for  $i$  if it's propensity score is nearest to that of  $i$

## NN Matching: Bias vs. Variance

- But how far away on the propensity score will you use? Herein lies the different types of matching proposed
  - Matching just one nearest neighbor minimizes bias at the cost of larger variance
  - Matching using additional nearest neighbors increases the bias but decreases the variance
- Matching with or without replacement
  - with replacement keeps bias low at the cost of larger variance
  - without replacement keeps variance low but at the cost of potential bias

## Distance between treatment and control units

- What was historically done was limiting “distance” through various *ad hoc* choices
- Imagine these choices as creating like a lasso (like the cowboy rope)
- Anything within the lasso could be used for the imputation; anything outside the lasso could not
- There were two common ways – caliper matching and radius matching.

## Caliper matching

- Caliper matching is a variation on NN matching that tries to build brakes into the algorithm as to avoid “bad neighbors”
- It does this by imposing a tolerable maximum distance (e.g., 0.2 units in the propensity score away from a treatment unit  $i$ 's propensity score)
- Note – this is a one-to-one imputation, and if there doesn't exist anybody in the control group unit  $j$  within that “caliper”, then treatment unit  $i$  is discarded
- Means we aren't estimating the ATE anymore once we start dropping units
- It's difficult to know what this caliper should be *ex ante*, hence why I said it is somewhat *ad hoc*

## Radius matching

- Each treatment unit  $i$  is matched with the control group units whose propensity score are in a predefined neighborhood of the propensity score of the treatment unit.
- **All** the control units with  $\hat{\rho}_j$  falling within a radius  $r$  from  $\hat{\rho}_i$  are matched to the treatment unit  $i$  – this is what distinguishes it from calipers, and makes it more similar to covariate matching (Abadie and Imbens 2006, 2008)
- The smaller the radius, the better the quality of the matches, but the higher the possibility some treatment units are not matched because the neighborhood does not contain control group units  $j$

# Software

- I think you can use `-teffects`, `psmatch`- to get at these two nearest neighbor approaches by setting the number of matches
- You can use `-pscore2-` for stratification, but I think the standard errors are wrong, so you may need to just do it manually using bootstrapping or variance approximation, and that may be a pain to program up
- Not sure of the R command, but I know it's out there

## King and Nielsen (2019)

- There is a King and Nielsen (2019) critique of these methods that is popularly known but not popularly studied
- King and Nielsen (2019) is not a critique of the propensity score, because it does not apply to stratification, regression adjustment, or inverse probability weighting
- It only applies to nearest neighbor and is related to forced balance through trimming and a myriad of other common choices made by the researcher

*“[The] more balanced the data, or the more balance it becomes by [trimming] some of the observations through matching, the more likely propensity score matching will degrade inferences.” – King and Nielsen (2019)*

# Examples of propensity score matching

- Workhorse example of propensity score matching is the Job Trainings Program (NSW)
- First studied by LaLonde (1986) evaluating multiple econometric models for program evaluation
  - All the standard estimators failed to estimate the known ATE when replacing experimental controls with non-experimental controls – even difference-in-differences
- Dehejia and Wahba (1999; 2002) use LaLonde's data with propensity score matching and found better results
- Critiques by Petra Todd, Jeff Smith and others followed which I won't review here for sake of time

# Description of NSW Job Trainings Program

The National Supported Work Demonstration (NSW), operated by Manpower Demonstration Research Corp in the mid-1970s:

- was a temporary employment program designed to help disadvantaged workers lacking basic job skills move into the labor market by giving them work experience and counseling in a sheltered environment
- was also unique in that it **randomly assigned** qualified applicants to training positions:
  - **Treatment group**: received all the benefits of NSW program
  - **Control group**: left to fend for themselves
- admitted AFDC females, ex-drug addicts, ex-criminal offenders, and high school dropouts of both sexes

# NSW Program

- Treatment group members were:
  - guaranteed a job for 9-18 months depending on the target group and site
  - divided into crews of 3-5 participants who worked together and met frequently with an NSW counselor to discuss grievances and performance
  - paid for their work
- Control group members were randomized so the same
- Note: the randomization balanced observables and unobservables across the two arms, thus enabling the estimation of an ATE for the people who self-selected into the program

# NSW Program

- Other details about the NSW program:
  - Wages: NSW offered the trainees lower wage rates than they would've received on a regular job, but allowed their earnings to increase for satisfactory performance and attendance
  - Post-treatment: after their term expired, they were forced to find regular employment
  - Job types: varied within sites – gas station attendant, working at a printer shop – and males and females were frequently performing different kinds of work

# NSW Data

- NSW data collection:
  - MDRC collected earnings and demographic information from both treatment and control at baseline and every 9 months thereafter
  - Conducted up to 4 post-baseline interviews
  - Different sample sizes from study to study can be confusing, but has simple explanations

# NSW Data

- Estimation:
  - NSW was a randomized job trainings program; therefore estimating the average treatment effect is straightforward:

$$\frac{1}{N_t} \sum_{D_i=1} Y_i - \frac{1}{N_c} \sum_{D_i=0} Y_i \approx E[Y^1 - Y^0]$$

in large samples assuming treatment selection is independent of potential outcomes (randomization) – i.e.,  $(Y^0, Y^1) \perp\!\!\!\perp D$ .

- NSW worked: Treatment group participants' real earnings post-treatment (1978) was positive and economically meaningful –  
 $\approx \$900$  (LaLonde 1986) to  $\$1,800$  (Dehejia and Wahba 2002)  
depending on the sample used

LaLonde, Robert J. (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data". *American Economic Review*.

LaLonde's study was **not** an evaluation of the NSW program, as that had been done, but rather an evaluation of econometric models done by:

- replacing the experimental NSW control group with non-experimental control group drawn from two nationally representative survey datasets: Current Population Survey (CPS) and Panel Study of Income Dynamics (PSID)
- estimating the average effect using non-experimental workers as controls for the NSW trainees
- comparing his non-experimental estimates to the experimental estimates of \$900

## LaLonde (1986)

- LaLonde's conclusion: available econometric approaches were biased and inconsistent
  - His estimates were way off and usually the wrong sign
  - Conclusion was influential in policy circles and led to greater push for more experimental evaluations

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW  
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA<sup>a,b</sup>

Name of Comparison Group <sup>d</sup>	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted <sup>c</sup> (3)	Unadjusted (4)	Adjusted <sup>c</sup> (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted <sup>c</sup> (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	\$455 (704)	\$697 (760)	-\$509 (967)	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

<sup>a</sup> The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

<sup>b</sup> Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

<sup>c</sup> The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

<sup>d</sup> See Table 3 for definitions of the comparison groups.

TABLE 5—EARNINGS COMPARISONS AND ESTIMATED TRAINING EFFECTS FOR THE NSW  
MALE PARTICIPANTS USING COMPARISON GROUPS FROM THE PSID AND THE CPS-SSA<sup>a,b</sup>

Name of Comparison Group <sup>d</sup>	Comparison Group Earnings Growth 1975–78 (1)	NSW Treatment Earnings Less Comparison Group Earnings				Difference in Differences: Difference in Earnings		Unrestricted Difference in Differences:		Controlling for All Observed Variables and Pre-Training Earnings (10)	
		Pre-Training Year, 1975		Post-Training Year, 1978		Growth 1975–78 Treatments Less Comparisons		Quasi Difference in Earnings Growth 1975–78			
		Unadjusted (2)	Adjusted <sup>c</sup> (3)	Unadjusted (4)	Adjusted <sup>c</sup> (5)	Without Age (6)	With Age (7)	Unadjusted (8)	Adjusted <sup>c</sup> (9)		
Controls	\$2,063 (325)	\$39 (383)	\$-21 (378)	\$886 (476)	\$798 (472)	\$847 (560)	\$856 (558)	\$897 (467)	\$802 (467)	\$662 (506)	
PSID-1	\$2,043 (237)	-\$15,997 (795)	-\$7,624 (851)	-\$15,578 (913)	-\$8,067 (990)	\$425 (650)	-\$749 (692)	-\$2,380 (680)	-\$2,119 (746)	-\$1,228 (896)	
PSID-2	\$6,071 (637)	-\$4,503 (608)	-\$3,669 (757)	-\$4,020 (781)	-\$3,482 (935)	\$484 (738)	-\$650 (850)	-\$1,364 (729)	-\$1,694 (878)	-\$792 (1024)	
PSID-3	(\$3,322 (780))	(\$455 (539))	(\$455 (704))	(\$697 (760))	(\$509 (967))	\$242 (884)	-\$1,325 (1078)	\$629 (757)	-\$552 (967)	\$397 (1103)	
CPS-SSA-1	\$1,196 (61)	-\$10,585 (539)	-\$4,654 (509)	-\$8,870 (562)	-\$4,416 (557)	\$1,714 (452)	\$195 (441)	-\$1,543 (426)	-\$1,102 (450)	-\$805 (484)	
CPS-SSA-2	\$2,684 (229)	-\$4,321 (450)	-\$1,824 (535)	-\$4,095 (537)	-\$1,675 (672)	\$226 (539)	-\$488 (530)	-\$1,850 (497)	-\$782 (621)	-\$319 (761)	
CPS-SSA-3	\$4,548 (409)	\$337 (343)	\$878 (447)	-\$1,300 (590)	\$224 (766)	-\$1,637 (631)	-\$1,388 (655)	-\$1,396 (582)	\$17 (761)	\$1,466 (984)	

<sup>a</sup> The columns above present the estimated training effect for each econometric model and comparison group. The dependent variable is earnings in 1978. Based on the experimental data an unbiased estimate of the impact of training presented in col. 4 is \$886. The first three columns present the difference between each comparison group's 1975 and 1978 earnings and the difference between the pre-training earnings of each comparison group and the NSW treatments.

<sup>b</sup> Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

<sup>c</sup> The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high school dropout status, and race.

<sup>d</sup> See Table 3 for definitions of the comparison groups.

# Imbalanced covariates for experimental and non-experimental samples

covariate	mean	(s.d.)	CPS	NSW	t-stat	diff
			All	Controls		
			$N_c = 15,992$	$N_t = 297$		
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-0.43
Education	12.0	2.86	12.03	10.38	9.85	1.65
1975 Earnings	13.51	9.31	13.65	3.1	19.63	10.6
1975 Unemp	0.11	0.32	0.11	0.37	14.29	-0.26

## Dehejia and Wahba (1999)

- Dehejia and Wahba (DW) update LaLonde's original study using propensity score matching
  1. Dehejia, Rajeev H. and Sadek Wahba (1999). "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs". Journal of the American Statistical Association, vol. 94(448): 1053-1062 (pdf)
- Can propensity score matching improve over the estimators that LaLonde examined?

Table 1. Sample Means of Characteristics for NSW and Comparison Samples

	No. of observations	Age	Education	Black	Hispanic	No degree	Married	RE74 (U.S. \$)	RE75 (U.S. \$)
<b>NSW/Lalonde:<sup>a</sup></b>									
Treated	297	24.63 (.32)	10.38 (.09)	.80 (.02)	.09 (.01)	.73 (.02)	.17 (.02)	3,066 (236)	
Control	425	24.45 (.32)	10.19 (.08)	.80 (.02)	.11 (.02)	.81 (.02)	.16 (.02)	3,026 (252)	
<b>RE74 subset:<sup>b</sup></b>									
Treated	185	25.81 (.35)	10.35 (.10)	.84 (.02)	.059 (.01)	.71 (.02)	.19 (.02)	2,096 (237)	1,532 (156)
Control	260	25.05 (.34)	10.09 (.08)	.83 (.02)	.1 (.02)	.83 (.02)	.15 (.02)	2,107 (276)	1,267 (151)
<b>Comparison groups:<sup>c</sup></b>									
PSID-1	2,490	34.85 [.78]	12.11 [.23]	.25 [.03]	.032 [.01]	.31 [.04]	.87 [.03]	19,429 [991]	19,063 [1,002]
PSID-2	253	36.10 [1.00]	10.77 [.27]	.39 [.04]	.067 [.02]	.49 [.05]	.74 [.04]	11,027 [853]	7,569 [695]
PSID-3	128	38.25 [1.17]	10.30 [.29]	.45 [.05]	.18 [.03]	.51 [.05]	.70 [.05]	5,566 [686]	2,611 [499]
CPS-1	15,992	33.22 [.81]	12.02 [.21]	.07 [.02]	.07 [.02]	.29 [.03]	.71 [.03]	14,016 [705]	13,650 [682]
CPS-2	2,369	28.25 [.87]	11.24 [.19]	.11 [.02]	.08 [.02]	.45 [.04]	.46 [.04]	8,728 [667]	7,397 [600]
CPS-3	429	28.03 [.87]	10.23 [.23]	.21 [.03]	.14 [.03]	.60 [.04]	.51 [.04]	5,619 [552]	2,467 [288]

NOTE: Standard errors are in parentheses. Standard error on difference in means with RE74 subset/treated is given in brackets. Age = age in years; Education = number of years of schooling; Black = 1 if black, 0 otherwise; Hispanic = 1 if Hispanic, 0 otherwise; No degree = 1 if no high school degree, 0 otherwise; Married = 1 if married, 0 otherwise; RE74 = earnings in calendar year 19x.

<sup>a</sup> NSW sample as constructed by Lalonde (1986).

<sup>b</sup> The subset of the Lalonde sample for which RE74 is available.

<sup>c</sup> Definition of comparison groups (Lalonde 1986):

PSID-1: All male household heads under age 55 who did not classify themselves as retired in 1975.

PSID-2: Selects from PSID-1 all men who were not working when surveyed in the spring of 1976.

PSID-3: Selects from PSID-2 all men who were not working in 1975.

CPS-1: All CPS males under age 55.

CPS-2: Selects from CPS-1 all males who were not working when surveyed in March 1976.

CPS-3: Selects from CPS-2 all the unemployed males in 1976 whose income in 1975 was below the poverty level.

PSID-1 and CPS-1 are identical to those used by Lalonde. CPS2-3 are similar to those used by Lalonde, but Lalonde's original subset could not be recreated.

Table 2. Lalonde's Earnings Comparisons and Estimated Training Effects for the NSW Male Participants Using Comparison Groups From the PSID and the CPS<sup>a</sup>

A. Lalonde's original sample										B. RE74 subsample (results do not use RE74)								C. RE74 subsample (results use RE74)														
Comparison group	NSW				NSW				NSW				NSW				NSW				NSW											
	treatment	Unrestricted differences	treatment	Unrestricted differences in	earnings less comparison	Quasi-difference	treatment	Unrestricted differences	earnings less comparison	Quasi-difference	treatment	Unrestricted differences in	earnings less comparison	Quasi-difference	treatment	Unrestricted differences	earnings less comparison	Quasi-difference	treatment	Unrestricted differences in	earnings less comparison	Quasi-difference	treatment	Unrestricted differences in	earnings less comparison	Quasi-difference						
	earnings	group	earnings	growth	group	earnings	earnings	group	earnings	growth	earnings	group	earnings	growth	earnings	group	earnings	growth	earnings	group	earnings	growth	earnings	group	earnings	growth						
	1978		1975–1978				1978		1978		1978		1978		1978		1978		1978		1978		1978		1978							
	Controlling for all variables		Controlling for all variables				Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables		Controlling for all variables							
	Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>	Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>		Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>		Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>		Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>		Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>		Unadjusted <sup>b</sup>	Adjusted <sup>c</sup>		Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>		Unadjusted <sup>d</sup>	Adjusted <sup>e</sup>							
	(1)	(2)	(3)	(4)		(1)	(2)		(3)	(4)		(1)	(2)		(3)	(4)		(1)	(2)		(3)	(4)		(3)	(4)		(5)					
NSW	886	798	879	802	820	1,794	1,672	1,750	1,631	1,612	1,794	1,668	1,750	1,672	1,655	1,750	1,672	1,655	1,750	1,672	1,655	1,750	1,672	1,655	1,750	1,672						
	(472)	(472)	(467)	(468)	(468)	(633)	(637)	(632)	(637)	(637)	(633)	(636)	(632)	(638)	(640)																	
PSID-1	-15,578	-8,057	-2,380	-2,119	-1,844	-15,205	-7,741	-582	-265	186	-15,205	-879	-582	-218	731																	
	(913)	(990)	(680)	(746)	(762)	(1155)	(1175)	(841)	(861)	(901)	(1155)	(931)	(841)	(866)	(886)																	
PSID-2	-4,020	-3,482	-1,364	-1,694	-1,876	-3,647	-2,810	721	298	111	-3,647	94	721	907	688																	
	(781)	(935)	(729)	(878)	(885)	(960)	(1082)	(886)	(1004)	(1032)	(960)	(1042)	(886)	(1004)	(1028)																	
PSID-3	697	-509	629	-552	-576	1,070	35	1,370	243	298	1,070	821	1,370	822	825																	
	(760)	(967)	(757)	(967)	(968)	(900)	(1101)	(897)	(1101)	(1105)	(900)	(1100)	(897)	(1101)	(1104)																	
CPS-1	-8,870	-4,416	-1,543	-1,102	-987	-8,498	-4,417	-78	525	709	-8,498	-8	709	-78	739	972																
	(562)	(577)	(426)	(450)	(452)	(712)	(714)	(537)	(567)	(560)	(712)	(572)	(560)	(572)	(557)	(557)																
CPS-2	-4,195	-2,341	-1,649	-1,129	-1,149	-3,822	-2,208	-263	371	305	-3,822	615	371	263	879	790																
	(533)	(620)	(459)	(551)	(551)	(671)	(746)	(574)	(662)	(666)	(671)	(672)	(666)	(672)	(574)	(574)	(574)															
CPS-3	-1,008	-1	-1,204	-263	-234	-635	375	-91	844	875	-635	1,270	844	875	-91	1,326	1,326															
	(539)	(681)	(532)	(677)	(675)	(657)	(821)	(641)	(808)	(810)	(657)	(798)	(641)	(798)	(798)	(798)	(798)															

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID-1, and CPS-1 are identical to Lalonde's. CPS-2 and CPS-3 are similar but not identical, because we could not exactly recreate his subset. Column (5) differs because the data file that we obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

<sup>a</sup> Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

<sup>b</sup> The estimates based on the NSW control group are unbiased estimates of the treatment impacts for the original sample (\$886) and for the RE74 sample (\$1,794).

<sup>c</sup> The exogenous variables used in the regressions-adjusted equations are age, age squared, years of schooling, high school dropout status, and race (and RE74 in Panel C).

<sup>d</sup> Regresses RE78 on a treatment indicator and RE75.

<sup>e</sup> The same as (d), but controls for the additional variables listed under (c).

<sup>f</sup> Controls for all pretreatment covariates.

## Proposition 2

$$X \perp\!\!\!\perp D|p(X)$$

- Conditional on the propensity score, the covariates are independent of the treatment, suggesting that the distribution of covariate values should be the same for both treatment and control groups
- This can be checked as we have data on all three once we've estimated the propensity score

## Trimming the data

- Common terms are “trimming” or “pruning”
- Drop units which do not overlap in terms of estimated propensity score
- Sometimes as a rule of thumb, just keep units on the  $[0.1, 0.9]$  interval

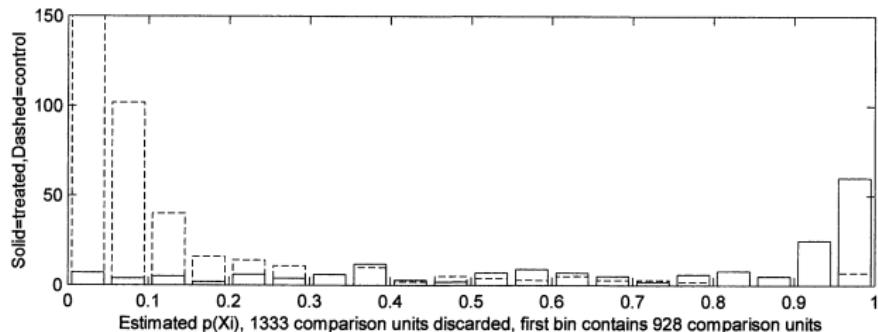


Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 928 PSID units. There is minimal overlap between the two groups. Three bins (.8-.85, .85-.9, and .9-.95) contain no comparison units. There are 97 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

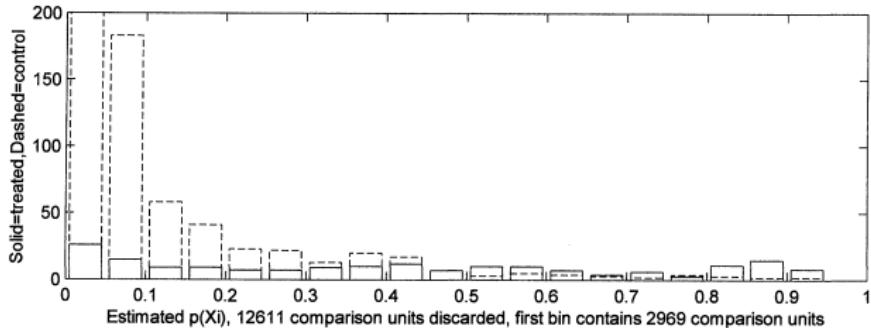


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units whose estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45-.5) contains no comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

Table 3. Estimated Training Effects for the NSW Male Participants Using Comparison Groups From PSID and CPS

	<i>NSW earnings less comparison group earnings</i>		<i>NSW treatment earnings less comparison group earnings, conditional on the estimated propensity score</i>					
			<i>Stratifying on the score</i>			<i>Matching on the score</i>		
	<i>(1) Unadjusted</i>	<i>(2) Adjusted<sup>a</sup></i>	<i>(3)</i>	<i>(4) Unadjusted</i>	<i>(5) Adjusted</i>	<i>(6) Observations<sup>c</sup></i>	<i>(7) Unadjusted</i>	<i>(8) Adjusted<sup>d</sup></i>
NSW	1,794 (633)	1,672 (638)						
PSID-1 <sup>e</sup>	-15,205 (1,154)	731 (886)	294 (1,389)	1,608 (1,571)	1,494 (1,581)	1,255	1,691 (2,209)	1,473 (809)
PSID-2 <sup>f</sup>	-3,647 (959)	683 (1,028)	496 (1,193)	2,220 (1,768)	2,235 (1,793)	389	1,455 (2,303)	1,480 (808)
PSID-3 <sup>f</sup>	1,069 (899)	825 (1,104)	647 (1,383)	2,321 (1,994)	1,870 (2,002)	247	2,120 (2,335)	1,549 (826)
CPS-1 <sup>g</sup>	-8,498 (712)	972 (550)	1,117 (747)	1,713 (1,115)	1,774 (1,152)	4,117	1,582 (1,069)	1,616 (751)
CPS-2 <sup>g</sup>	-3,822 (670)	790 (658)	505 (847)	1,543 (1,461)	1,622 (1,346)	1,493	1,788 (1,205)	1,563 (753)
CPS-3 <sup>g</sup>	-635 (657)	1,326 (798)	556 (951)	1,252 (1,617)	2,219 (2,082)	514	587 (1,496)	662 (776)

<sup>a</sup> Least squares regression: RE78 on a constant, a treatment indicator, age, age<sup>2</sup>, education, no degree, black, Hispanic, RE74, RE75.<sup>b</sup> Least squares regression of RE78 on a quadratic on the estimated propensity score and a treatment indicator, for observations used under stratification; see note (g).<sup>c</sup> Number of observations refers to the actual number of comparison and treatment units used for (3)–(5); namely, all treatment units and those comparison units whose estimated propensity score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.<sup>d</sup> Weighted least squares: treatment observations weighted as 1, and control observations weighted by the number of times they are matched to a treatment observation [same covariates as (a)].

Propensity scores are estimated using the logistic model, with specifications as follows:

<sup>e</sup> PSID-1: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, married, no degree, black, Hispanic, RE74, RE75, RE74<sup>2</sup>, RE75<sup>2</sup>, u74\*black).<sup>f</sup> PSID-2 and PSID-3: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, no degree, married, black, Hispanic, RE74, RE74<sup>2</sup>, RE75, RE75<sup>2</sup>, u74, u75).<sup>g</sup> CPS-1, CPS-2, and CPS-3: Prob ( $T_i = 1$ ) = F(age, age<sup>2</sup>, education, education<sup>2</sup>, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education\*RE74, age<sup>3</sup>).

*Table 4. Sample Means of Characteristics for Matched Control Samples*

<i>Matched samples</i>	<i>No. of observations</i>	<i>Age</i>	<i>Education</i>	<i>Black</i>	<i>Hispanic</i>	<i>No degree</i>	<i>Married</i>	<i>RE74 (U.S. \$)</i>	<i>RE75 (U.S. \$)</i>
NSW	185	25.81	10.35	.84	.06	.71	.19	2,096	1,532
MPSID-1	56	26.39	10.62	.86	.02	.55	.15	1,794	1,126
		[2.56]	[.63]	[.13]	[.06]	[.13]	[.12]	[1,406]	[1,146]
MPSID-2	49	25.32	11.10	.89	.02	.57	.19	1,599	2,225
		[2.63]	[.83]	[.14]	[.08]	[.16]	[.16]	[1,905]	[1,228]
MPSID-3	30	26.86	10.96	.91	.01	.52	.25	1,386	1,863
		[2.97]	[.84]	[.13]	[.08]	[.16]	[.16]	[1,680]	[1,494]
MCPS-1	119	26.91	10.52	.86	.04	.64	.19	2,110	1,396
		[1.25]	[.32]	[.06]	[.04]	[.07]	[.06]	[841]	[563]
MCPS-2	87	26.21	10.21	.85	.04	.68	.20	1,758	1,204
		[1.43]	[.37]	[.08]	[.05]	[.09]	.08	[896]	[661]
MCPS-3	63	25.94	10.69	.87	.06	.53	.13	2,709	1,587
		[1.68]	[.48]	[.09]	[.06]	[.10]	[.09]	[1,285]	[760]

NOTE: Standard error on the difference in means with NSW sample is given in brackets.

MPSID1-3 and MCPS1-3 are the subsamples of PSID1-3 and CPS1-3 that are matched to the treatment group.

## Subsequent studies

- Heckman et al. (1996, 1998) used experimental data from the US National Job Training Partnership Act (JTPA)
- They conclude that in order for matching estimators to have low bias, it is important that the data include a rich set of variables related to program participation and labor market outcomes, that the nonexperimental comparison group be drawn from the same local labor markets as the participants and the dependent variable (typically earnings) be measured in the same way for participants and nonparticipants
- All three of these conditions fail to hold in DW (1999, 2002) according to Smith and Todd (2005)

## Smith and Todd

- Difference-in-differences with propensity scores tended to work well in Smith and Todd (2005)
- But hard to make this a rule, because it's hard to know ex ante if we've specified the propensity score correctly (i.e., have CIA)
- It is vital you know your data, if you're going to use these methods, which means understanding at a deep level the way in which selection (i.e., treatment assignment) works in your data

# Coarsened exact matching

- There are two kinds of matching as we've said
  1. *Exact matching* matches a treated unit to all of the control units with the same covariate value. Sometimes this is impossible (e.g., continuous covariate).
  2. *Approximate matching* specifies a metric to find control units that are close to the treated unit. Requires a distance metric, such as Euclidean, Mahalanobis, or the propensity score. All of which can be implemented in Stata's `teffects`.
- Iacus, King and Porro (2011) propose another version of matching they call coarsened exact matching (CEM). Some big picture ideas

## Checking imbalance

- Iacus, King and Porro (2008) say that in practice approximate matching requires setting the matching solution beforehand, then checking for imbalance after.
- Start over, repeat, until the user is exhausted by checking for imbalance.

# CEM Algorithm

1. Begin with covariates  $X$ . Make a copy called  $X^*$
2. Coarsen  $X^*$  according to user-defined cutpoints or CEM's automatic binning algorithm
  - Schooling → less than high school, high school, some college, college, post college
3. Create one stratum per unique observation of  $X^*$  and place each observation in a stratum
4. Assign these strata to the original data,  $X$ , and drop any observation whose stratum doesn't contain at least one treated and control unit

You then add weights for stratum size and analyze without matching.

## Tradeoffs

- Larger bins mean more coarsening. This results in fewer strata.
- Fewer strata result in more diverse observations within the same strata and thus higher imbalance
- CEM prunes both treatment and control group units, which changes the parameter of interest. Be transparent about this as you're not estimating the ATE or the ATT when you start pruning

## Benefits

- The key benefit of CEM is that it is in a class of matching methods called *monotonic imbalance bounding*
- MIB methods bound the maximum imbalance in some feature of the empirical distributions by an ex ante decision by the user
- In CEM, this ex ante choice is the coarsening decision
- By choosing the coarsening beforehand, users can control the amount of imbalance in the matching solution
- It's also wicked fast.

## Imbalance

- There are several ways of measuring imbalance, but here we focus on the  $\mathcal{L}_1(f, g)$  measure which is

$$\mathcal{L}_1(f, g) = \frac{1}{2} \sum_{l_1 \dots l_k} |f_{l_1 \dots l_k} - g_{l_1 \dots l_k}|$$

where the  $f$  and  $g$  record the relative frequencies for the treatment and control group units.

- Perfect global imbalance is indicated by  $\mathcal{L}_1 = 0$ . Larger values indicate larger imbalance between the groups, with a maximum of  $\mathcal{L}_1 = 1$ .

# Stata

- Download `cem` from Stata: `ssc install cem, replace`
- You will automatically compute the global imbalance measure, as well as several unidimensional measures of imbalance, when using `cem`
- I got a  $\mathcal{L}_1 = 0.55$ . What does it mean?
  - By itself, it's meaningless. It's a reference point between matching solutions.
  - Once we have a matching solution, we will compare its  $\mathcal{L}_1$  to 0.55 and gauge the increase in balance due to the matching solution from that difference.
  - Thus  $\mathcal{L}_1$  works for imbalance as  $R^2$  works for model fit: the absolute values mean less than comparisons between matching solutions.

## More Stata

- Because `cem` bounds the imbalance *ex ante*, the most important information in the Stata output is the number of observations matched.
- You can also choose the coarsening as opposed to relying on the algorithm's automated binning.
- Once you have estimated the strata, you regress the outcome onto the treatment and then weight the regression by `cem_weights`. For instance,

```
regress re78 treat [iweight=cem_weight]
```

- For more on this, see Blackwell, et al. Stata journal article from 2009.