
MIXTAPE
SESSIONS

Roadmap

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:

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

But for every i unit, who is either treated or not treated, we are missing one of these (i.e., counterfactuals), and so we will *estimate* the missing counterfactual for one group with the observed values of the other

Defined causal parameters and weights

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

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

- w_{ij} is the weight unit j receives in predicting unit i's counterfactual
- Weighting is therefore an important part of what we will be doing, even though what we are going to be doing always is imputing missing counterfactuals through estimation

Estimating missing counterfactuals

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

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

What is selection on observables?

- Selection on observables is an estimation strategy that attempts to estimate average treatment effects using covariate adjustments that meet the backdoor criterion
- Very strong condition: conditional on covariates, all remaining variation in the treatment is “as good as randomly assigned”
- Covariates are used to construct a credible counterfactual as a weighted average of the other treatment category group (“look-a-like”)
- Estimators differ in how the weights, w_{ij}, are calculated and how comparisons are made

DAGs are crucial for selection on observables

- If you are going to estimate causal effects using covariates, then you **must use a model** as you will be assuming that you have not forgotten a confounder or included a collider
- You must **defend** selection on observables methods and it's much less credible to many people because they cannot confirm much of the assumptions
- Personal observation: given the lack of comfort people have with modeling in general, it's not a great sign when you see them then use selection on observables

Selection on 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)

Regressions that condition on covariates are with binary treatments can be used also

Exogenous Covariate

Definition of a Exogenous Covariate

Variable X is an exogenous covariate if for each individual i, the value of X_i does not vary because of the treatment status (i.e., not a collider)

- Does not imply X and treatment status are independent; just means that the treatment doesn't cause X (i.e., colliders aren't covariates)
- Covariates can be time invariant or change over the time – that's not relevant
- Goal is to use covariates that capture all confounders (no unobserved ones) and don't open backdoors through bad controls (e.g., colliders)

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

- 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

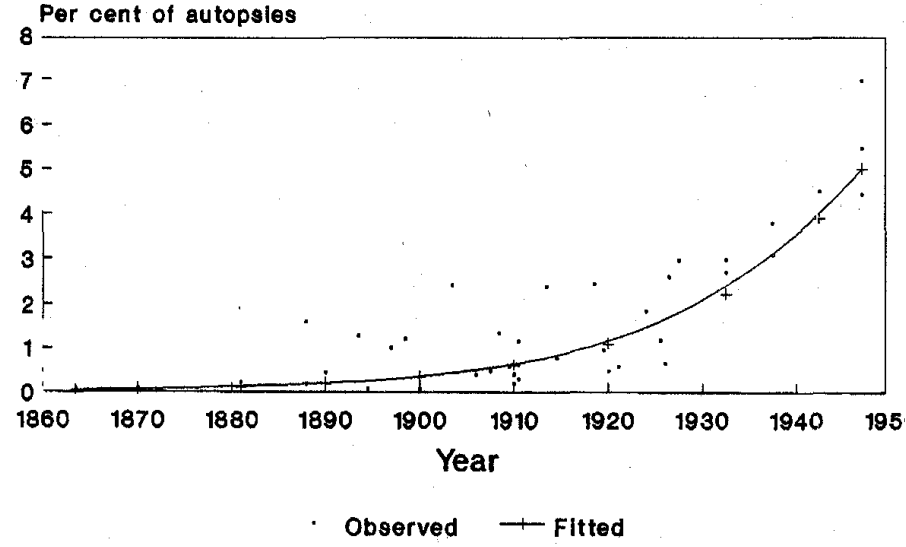


Figure 2(a)
Mortality from Cancer of the Lung in Males

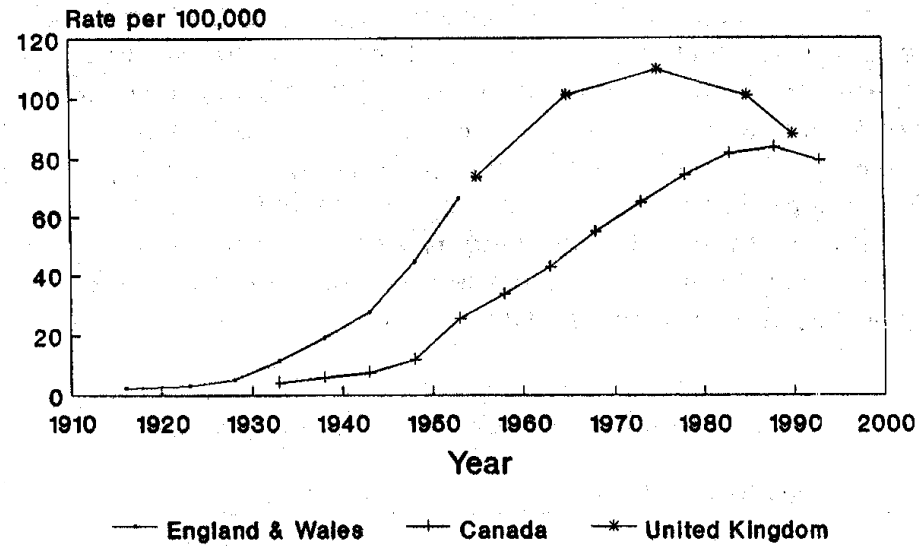
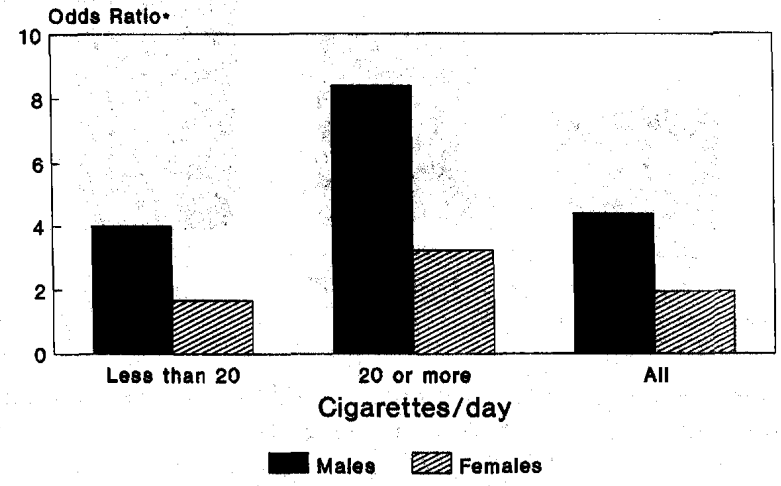
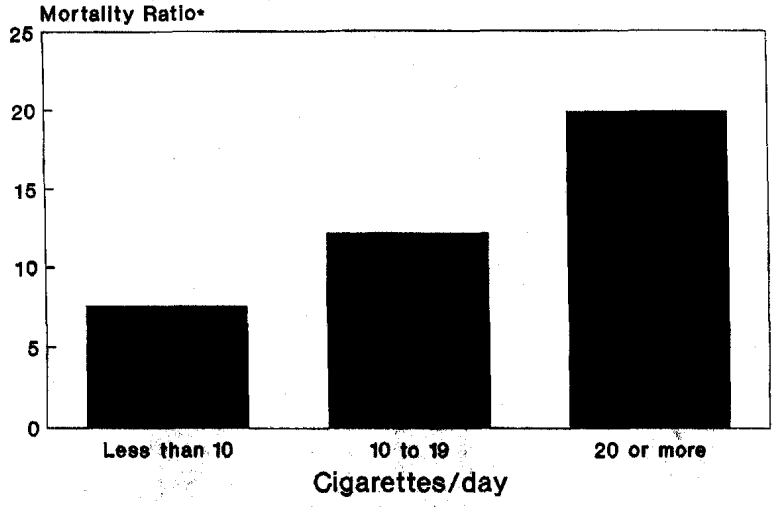


Figure 4
Smoking and Lung Cancer Case-control Studies



* Weighted mean

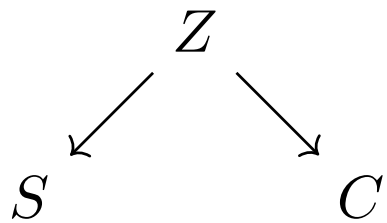
Figure 5
Smoking and Lung cancer Cohort Studies in Males



* weighted mean of 4 studies

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 and adjusting for *observable* differences became a solution

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?

Covariate imbalance

- Covariates are *not balanced* – their mean values differ for treatment and control group.
- We will use weighting and imputation to achieve covariate balance
- Keeping in mind we need this to be done for *relevant* covariates

General principles

Comparing Within Covariate Strata

- Create slices of the covariates called “strata”
- Compare mortality rates across the different smoking groups but *within* each strata
- Compare within covariate strata and then combine differences to neutralize observed confounders

General principles

Weighting

- Weight the data so that covariates are balanced,
- Then compare mortality across treatment and control
- Focus then is how to find those weights that cause covariate balance

Subclassification/stratification

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 always need three things to estimate a causal effect

- **Correct Assumptions**: what must we assume is true so that our models work with data?
- **Properly Measured 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!)

Recall the RCT Assumption of 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, but we could just as easily insert backdoor criterion
- Sometimes this is called also "unconfoundedness"

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

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

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

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

Summarizing

Weighted averages under either assumption:

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

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

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

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

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

Weighting by Age ($K = 2$)

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

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

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

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

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

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATE}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N^k}{N} \right)$?

Not identified! What went wrong?

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

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

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

	Death Rate			Number of	
X_k	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 $\widehat{\delta_{ATT}} = \sum_{k=1}^{K} (\overline{Y}^{1,k} - \overline{Y}^{0,k}) \cdot \left(\frac{N_T^k}{N_T} \right)$?

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

Curse of Dimensionality

- Stratification methods, including OLS, 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 strata is 3^k. For $k = 10$, then it's $3^{10} = 59,049$
- The problem isn't just the number of covariates; it's the number of strata based on those covariates (you can hit the curse fast)

Curse of Dimensionality

- If sparseness occurs, it means many cells may contain either only treatment units or only control units but not both, and that violates our second assumption
- 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
- Simple weighting methods is also a problem if the cells are "too coarse"

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 δ_{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 δ_{ATE}. In that case, we match in both directions:
  1. If observation i is treated, we impute Y_i^0 using the control matches, $\{Y_{j_1(i)}, \ldots, Y_{j_M(i)}\}$
  2. If observation i is control, we impute Y_i^1 using the treatment matches, $\{Y_{j_1(i)}, \ldots, Y_{j_M(i)}\}$
- The estimator is:

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

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		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
Average:		

A Training Example

Trainees

unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees

unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample

unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
Average:		

A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:		

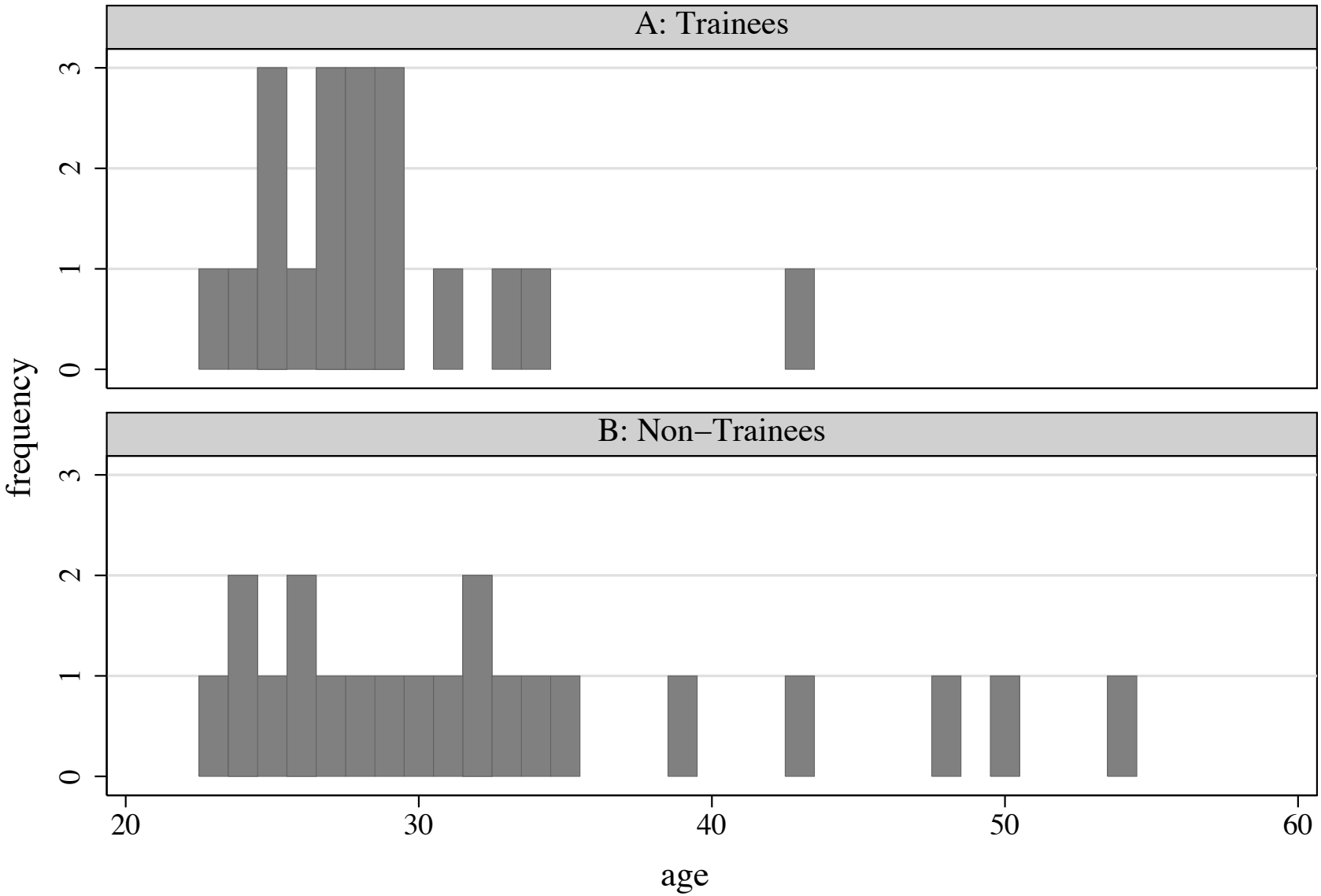
A Training Example

Trainees		
unit	age	earnings
1	28	17700
2	34	10200
3	29	14400
4	25	20800
5	29	6100
6	23	28600
7	33	21900
8	27	28800
9	31	20300
10	26	28100
11	25	9400
12	27	14300
13	29	12500
14	24	19700
15	25	10100
16	43	10700
17	28	11500
18	27	10700
19	28	16300
Average:	28.5	16426

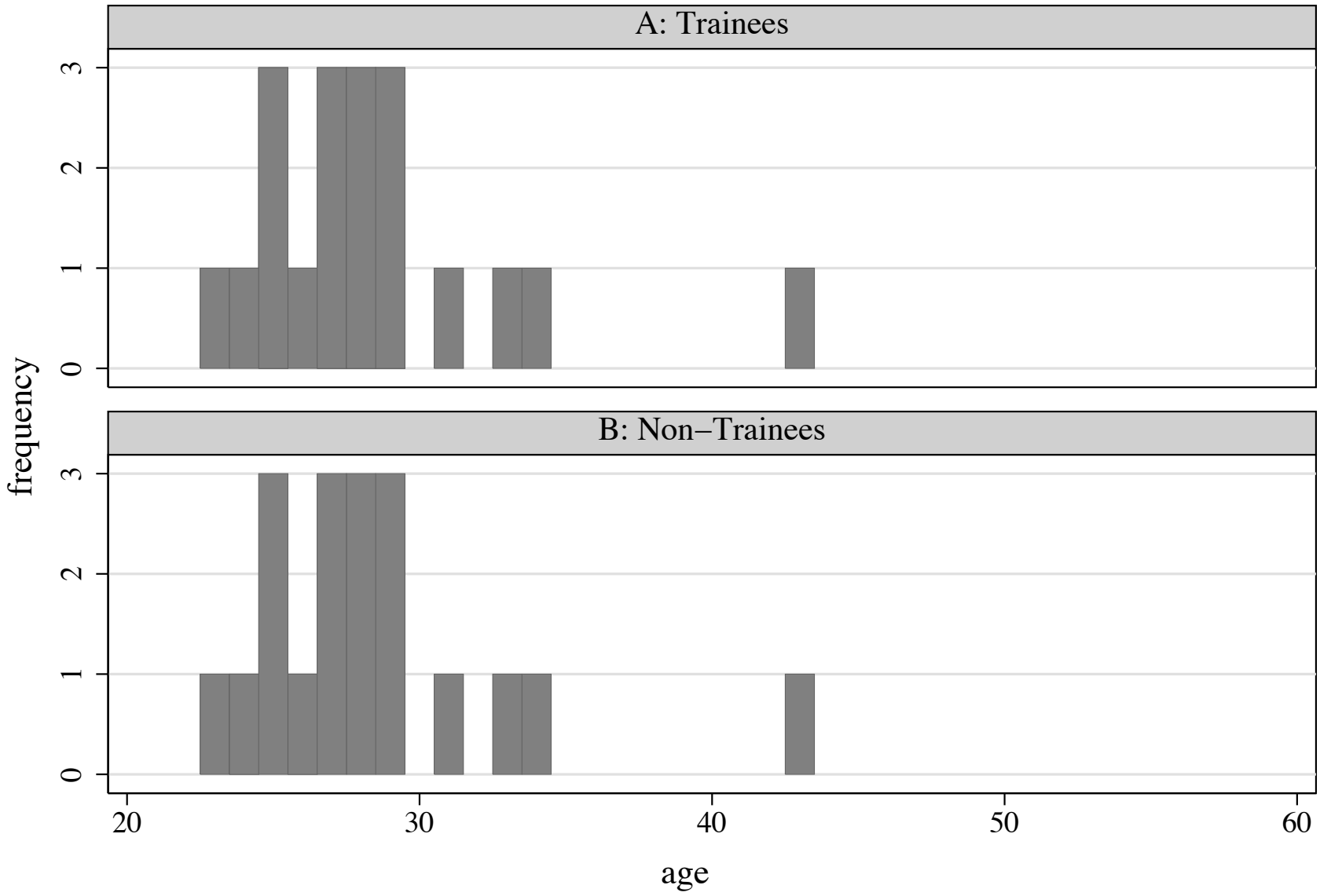
Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

Matched Sample		
unit	age	earnings
8	28	8800
14	34	24200
17	29	6200
15	25	23300
17	29	6200
20	23	9500
10	33	15500
4	27	9300
12	31	26600
11,13	26	8450
15	25	23300
4	27	9300
17	29	6200
9,16	24	17700
15	25	23300
1	43	20900
8	28	8800
4	27	9300
8	28	8800
Average:	28.5	13982

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:

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

where

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

Normalized Euclidean distance

- Notice that the normalized Euclidean distance is equal to:

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

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

Mahalanobis distance

Definition: Mahalanobis distance

The Mahalanobis distance is the scale-invariant distance metric:

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

where $\widehat{\Sigma}_X$ 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 ω_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

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

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.

$$\begin{aligned}
\mu^0(x) &= E[Y|X = x, D = 0] = E[Y^0|X = x], \\
\mu^1(x) &= E[Y|X = x, D = 1] = E[Y^1|X = x], \\
Y_i &= \mu^{D_i}(X_i) + \varepsilon_i
\end{aligned}$$

Deriving the matching bias

Substitute and distribute terms

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

Deriving the matching bias

Difference between sample estimate and population parameter is:

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

Algebraic manipulation and simplification:

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

Deriving the matching bias

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

Deriving the matching bias

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

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

Deriving the matching bias

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

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

Now consider the implications if k is large:

Deriving the matching bias

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

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

Now consider the implications if k is large:

- The difference between X_i and $X_{j(i)}$ converges to zero very slowly

Deriving the matching bias

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

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

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

Deriving the matching bias

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

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

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
- $E[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1]$ may not converge to zero and an be very large!

Deriving the matching bias

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

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

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

Deriving the matching bias

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

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

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
- $E[\sqrt{\frac{1}{N}}(\mu^0(X_i) - \mu^0(X_{j(i)}))|D = 1]$ may not converge to zero and an be very large!
- $E[\sqrt{\frac{1}{N}}(\widehat{\delta}_{ATT} - \delta_{ATT})]$ 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 $||X_i - X_{j(i)}||$ which is virtually guaranteed by the curse of dimensionality. 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:

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

  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 under Treatment	Potential Outcome under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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

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

Bias adjustment in matched data

unit	Potential Outcome under Treatment	Potential Outcome under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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 $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

Bias adjustment in matched data

	Potential Outcome			
unit	under Treatment	under Control		
i	Y_i^1	Y_i^0	D_i	X_i
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 $\widehat{\mu^0}(X) = \widehat{\beta}_0 + \widehat{\beta}_1 X = 2 + X$

$$\widehat{\delta}_{ATT} = \frac{(10-8) - (\widehat{\mu^0}(3) - \widehat{\mu^0}(4))}{3} + \frac{(4-1) - (\widehat{\mu^0}(1) - \widehat{\mu^0}(0))}{3}$$

$$+ \frac{(10-9) - (\widehat{\mu^0}(10) - \widehat{\mu^0}(8))}{3} = 1.33$$

Matching bias: Implications for practice

Matching bias arises because of the effect of large matching discrepancies on $\mu^0(X_i) - \mu^0(X_{j(i)})$ due to a lack of common support. 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):

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

- For matching without replacement, the "usual" variance estimator:

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

  is valid.

Large sample distribution for matching estimators

- For matching with replacement:

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

  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 $\widehat{var}(\varepsilon_i|X_i, D_i=0))$

Final note about bias adjustment

- Identifying assumptions with selection on observables are:
  1. Potential outcomes are distributed independent of treatment status ("conditional independence")
  2. Common support
- Matching discrepancies are due to violations of second assumption caused by curse of dimensionality – not the first
- Matching bias adjustments **do not** recover ATE or ATT if conditional independence fails as that implies omitting a possibly very important and influential confounder

Avoiding dimensionality problems

- Curse of dimensionality makes matching on K covariates challenging but earlier work before nearest neighbor covariate matching existed
- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method which reduced K covariates used for adjusting into a single scalar
- Insofar as treatment is random conditional on K covariates, then one can use the propensity score to adjust for confounders
- Variety of ways to incorporate the propensity score, but first we describe the propensity score as a dimension reduction method

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, some of which is not broadly appreciated, such as basic diagnostics provided by common support checks

Basic 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)
- Rather than compare units with similar values of X, we compare units with similar **estimated conditional probabilities of treatment**
- Important theorem shows that once we adjust comparisons using the propensity score, we do not 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: $p(X) = Pr(D = 1|X)$

Assumptions

Two sufficient and necessary identification assumptions:

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

With both, we can incorporate the propensity score into comparisons of treated and untreated units and obtain unbiased and consistent estimates of the ATE

Propensity score methods

Covariate adjustment using the propensity score is a three step process

1. Estimate the propensity score using logit/probit
2. Estimate a particular average treatment effect (e.g., ATE, ATT) incorporating the estimated propensity score (e.g., stratification, imputation, regression, or inverse probability weighting)
3. Estimate standard errors

Between steps 1 and 2 are various design-like diagnostic steps such as examining common support using histograms, trimming, etc.

Step 1: 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

$$\widehat{\rho}_i(X_i) = \widehat{\beta} X_i$$

- Note that each unit i now has a predicted probability of treatment given the values of their covariates relative to everyone else's
- Frequentist probability – you've basically just obtained the likelihood someone who "looks like you" would be treated (regardless of whether you were in fact treated)

Identification

- Write down the definition of the ATE conditional on X_i

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

- Given conditional independence, we can substitute average values of Y for potential outcomes using the switching equation:

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

  and similar for other term Y^0
- We need common support (assumption 2) so that both terms can be estimated

Propensity score theorem

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

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

Estimating ATE with propensity score

Unbiased Estimate of ATE

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

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

Propensity Score Theorem Proof

Details of the proof are provided for those who want to study it more closely

- First note 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.

- Rest of the proof is straightforward and I've drawn it out in case you need to see all the steps

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

Propensity score balances covariates

- D and X are independent conditional on $p(X)$:

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

- This implies that the distribution of the covariates should be the same for treatment and control groups:

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

- But we should check it ourselves. For propensity score ranges (e.g., 0.1 to 0.2, 0.2 to 0.3, ...), what percent of the treatment group are male? What percent of control group are male?

Checking common support

- Common support is required for unbiased estimation of the ATE or ATT, and it is often violated in practice depending on the distribution of the included confounders for selecting into treatment
- A histogram of propensity scores by treatment and control group is a key diagnostic in highlighting the overlap problem
- Crump, et al. (2009) suggest keeping units whose propensity scores are within the interval [0.1,0.9] (called “trimming”)
- Note that trimming comes at a price – you are no longer estimating the ATE or the ATT if you are dropping units

Estimation with the propensity score

- Propensity scores have many value to us, such as checking for common support, but the goal is ultimately to estimate an average treatment effect
- Many different ways have been developed over the years to incorporate the propensity score into estimation
- Sometimes you are imputing missing counterfactuals finding nearest neighbors with similar propensity scores, and sometimes you are weighting by the propensity score
- We'll discuss a few of them starting with weighting by the propensity score – inverse probability weighting (IPW)

Inverse probability weighting

- IPW uses the estimated propensity score to reweight the outcomes (e.g., Robins and Rotnitzky 1995, Imbens 2000, Hirano and Imbens 2001)
- The weights can be expressed in two ways (the difference being how well either approach can handle extreme values of the propensity score)
    1. Without normalization (Horvitz and Thompson 1952)
    2. Normalized (Hajek1971)
- We'll introduce IPW without normalization first as normalized weights can be a little intimidating at first glance

Inverse probability weighting

- IPW is non-parametric – you are just taking averages and multiplying by weights
- There are also fewer implementation choices – you aren't choosing how many neighbors to include, how far away a neighbor can be – but you still have to closely examine common support
- Fun fact: two new diff-in-diff estimators use IPW to incorporate covariates into estimation (Sant'anna and Zhao 2020; Callaway and Sant'anna 2020)

Inverse Probability Weighting

Estimating ATE with IPW

Given $Y^1, Y^0 \perp\!\!\!\perp D|X$ and common support, then

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

Inverse Probability Weighting

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 $P(X|D=1)$. □

Inverse Probability Weighting

Estimating ATT with IPW

Given $Y^0 \perp\!\!\!\perp D|X$ and common support, then

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

Similar proof as ATE

Weighting on the propensity score

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

1. Estimate ech unit i's propensity score: $\widehat{\rho}_i(X_i)$
2. Use estimated score to produce analog estimators. Let $\widehat{\delta}_{ATE}$ and $\widehat{\delta}_{ATT}$ be an estimate of the ATE and ATT parameter:

$$\widehat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\rho}_i(X_i)}{\widehat{\rho}_i(X_i)\cdot(1-\widehat{\rho}_i(X_i))}$$

$$\widehat{\delta}_{ATT} = \frac{1}{N_T}\sum_{i=1}^{N} Y_i \cdot \frac{D_i - \widehat{\rho}_i(X_i)}{1-\widehat{\rho}_i(X_i)}$$

Note that we are simply averaging and differencing after weighting each unit by its propensity score

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 $\rho(X)$
  - $\rightarrow$ Parameteric first step: Newey and McFadden (1994)
  - $\rightarrow$ Non-parametric first step: Newey (1994)
- IPW is a smooth estimator which means the bootstrap is valid for inference (Adudumilli 2018 and Bodory et al. 2020) unlike covariate nearest neighbor matching which Abadie and Imbens (2008) show is not valid

Implementation with software

- I like estimating with IPW manually because I like being reminded how simple a procedure it is
- You'll probably want to use Stata's -teffects- or R's -ipw- so that you can get standard errors
- Note that Stata's -teffects- uses the Hajek normalization weights which will produce identical estimates to my program in the Mixtape
- My book doesn't manually do the inference, but it would be fun and easy to do using the bootstrap

Double robust estimators

- Propensity scores and estimated propensity scores are not the same thing
- You can have the right covariates but the wrong model and unbiasedness requires the correct model
- What if you had a way to control for the covariates using propensity scores and something else like regression?
- Buys you some insurance against model misspecification if such a thing existed

Double robust estimators

- Lots of papers began to try and address the model misspecification problem by combining propensity scores with other methods called "double robust" (Robins and Rotnizky 1995; Hirano and Imbens 2001)
- Basic idea in all of them was to control for covariates twice *at the same time* without paying a price
- 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
- We give ourselves two chances to get it right (either/or not both/and) but if neither is properly specified, then you didn't really gain much

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 “imputation”, is another way that utilizes the $\widehat{\rho_i}(X_i)$
- Matching estimation based on the propensity score has the same first step as IPW, but not the second and third steps
- Common support starts to be more complex with imputation methods because you will need to decide how far away from a unit’s own propensity score is a tolerable distance to be considered a “neighbor”

Standard matching strategy

- Pair each treatment unit i with one or more *comparable* control group unit j, where comparability is in terms of proximity, or distance, 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

Let the ATT be our parameter of interest:

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

We estimate it as follows

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

where N_T is the number of matched treatment units in the sample. Note the difference between *imputation* and IPW – the only weight here is $\frac{1}{N_T}$

Matching methods

- The probability of observing two units with *exactly* the same propensity score is in principle zero if $Pr(X = x)$ 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 based on the propensity score is a multi step process that bears resemblance to the stratification/subclassification method proposed by Cochran (1968)
- Method uses brute force to achieve the balancing property discussed earlier, which is then used with weighted differences in means within propensity score “strata”
- Dehejia and Wahba (2002) used stratification matching in their seminal paper

Stratification: Achieving Balance

First create "propensity score strata" inside which you have balanced covariates

1. Sort the data by propensity score and divide into groups of observations with similar propensity scores (e.g., percentiles)
2. Within each strata, test (e.g., t-test) whether the means of the k covariates are equal between treatment and control
3. If so, then stop. If not, it means the covariates aren't balanced *within that propensity score strata* so then divide that strata in half and repeat step 2
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

Propensity score matching

- Next we review explicit imputation based on the propensity score or what is sometimes called propensity score matching
- King and Nielsen (2019) is a critique of using propensity scores *for matching* (i.e., imputation)
- But not a critique of the propensity score itself or to stratification, regression adjustment, or IPW
- Issues raised have to do with forced balance through trimming and a myriad of other common choices made by the researcher

Ad hoc user choices introduce bias

"'[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)

Nearest Neighbor

Pretty similar to covariate matching. Formula is

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

- N_T is the number of treated units i and N_C is number of control units j
- w_{ij} is equal to $\frac{1}{N_C}$ if j is a control unit and zero otherwise
- 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

How far away on the propensity score will you use is what makes some of the different types of matching proposed differ

- 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

NN Matching: Bias vs. 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 cowboy rope lasso that matches to everything inside that circle
- There were two common ways for creating the circle – 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" 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 which as with all trimming changes the parameter we are estimating
- 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 $\widehat{\rho_j}(X_j)$ falling within a radius r from $\widehat{\rho_i}(X_i)$ 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

- 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
- You can use the `MatchIt` package in R

Failure of econometric estimators (LaLonde 1986)

- Evaluation of the Job Trainings Program (NSW) has a rich history in causal inference
- Bob LaLonde (passed away November 2015) was a Card and Ashenfelter student at Princeton whose job market paper evaluated, not NSW itself, but econometric methods one would use in something like NSW
- Dehejia and Wahba (1999; 2002) used LaLonde's data with propensity score matching and found they could recover known effects
- Critiques by Petra Todd, Jeff Smith and others followed which I'll summarize

Summarizing LaLonde (1986)

- Very clever study that combined experimental and non-experimental data to ascertain whether popular econometric methods could recover unbiased effects when those effects were already known
- Damning conclusion – 1986 AER (it was LaLonde's JMP) found econometric methods failed to get the number right, and worse, failed to get the sign right
- Was a critical paper in the emerging "credibility crisis" within labor and helped fuel the type of work we now broadly consider to be design based causal inference

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

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:

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

    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

CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

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

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high sch dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

Switching out the control group

- Think of $800 to $900 as the “ground truth” since row 1 was using the RCT
- LaLonde “drops” the experimental controls (which satisfied independence) and “replaces” it with six different draws from two nationally representative surveys (PSID and CPS)
- Now the dataset contains a negatively selected treatment group compared to a nationally representative control group
- Will selection on observable methods “work”?

CPS-SSA-1	$1,196	−$10,585	−$4,654	−$8,870	−$4,416	$1,714	$195	−$1,543	−$1,102	−$805
	(61)	(539)	(509)	(562)	(557)	(452)	(441)	(426)	(450)	(484)
CPS-SSA-2	$2,684	−$4,321	−$1,824	−$4,095	−$1,675	$226	−$488	−$1,850	−$782	−$319
	(229)	(450)	(535)	(537)	(672)	(539)	(530)	(497)	(621)	(761)
CPS-SSA-3	$4,548	$337	$878	−$1,300	$224	−$1,637	−$1,388	−$1,396	$17	$1,466
	(409)	(343)	(447)	(590)	(766)	(631)	(655)	(582)	(761)	(984)

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

[b] Estimates are in 1982 dollars. The numbers in parentheses are the standard errors.

[c] The exogenous variables used in the regression adjusted equations are age, age squared, years of schooling, high sch dropout status, and race.

[d] See Table 3 for definitions of the comparison groups.

Imbalanced covariates for experimental and non-experimental samples

	All		CPS Controls $N_c = 15,992$	NSW Trainees $N_t = 297$		
covariate	mean	(s.d.)	mean	mean	t-stat	diff
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

Dehija 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. $)*
NSW/Lalonde:[a]									
Treated	297	24.63	10.38	.80	.09	.73	.17		3,066
		(.32)	(.09)	(.02)	(.01)	(.02)	(.02)		(236)
Control	425	24.45	10.19	.80	.11	.81	.16		3,026
		(.32)	(.08)	(.02)	(.02)	(.02)	(.02)		(252)
RE74 subset:[b]									
Treated	185	25.81	10.35	.84	.059	.71	.19	2,096	1,532
		(.35)	(.10)	(.02)	(.01)	(.02)	(.02)	(237)	(156)
Control	260	25.05	10.09	.83	.1	.83	.15	2,107	1,267
		(.34)	(.08)	(.02)	(.02)	(.02)	(.02)	(276)	(151)
Comparison groups:[c]									
PSID-1	2,490	34.85	12.11	.25	.032	.31	.87	19,429	19,063
		[.78]	[.23]	[.03]	[.01]	[.04]	[.03]	[991]	[1,002]
PSID-2	253	36.10	10.77	.39	.067	.49	.74	11,027	7,569
		[1.00]	[.27]	[.04]	[.02]	[.05]	[.04]	[853]	[695]
PSID-3	128	38.25	10.30	.45	.18	.51	.70	5,566	2,611
		[1.17]	[.29]	[.05]	[.03]	[.05]	[.05]	(686)	[499]
CPS-1	15,992	33.22	12.02	.07	.07	.29	.71	14,016	13,650
		[.81]	[.21]	[.02]	[.02]	[.03]	[.03]	[705]	[682]
CPS-2	2,369	28.25	11.24	.11	.08	.45	.46	8,728	7,397
		[.87]	[.19]	[.02]	[.02]	[.04]	[.04]	[667]	[600]
CPS-3	429	28.03	10.23	.21	.14	.60	.51	5,619	2,467
		[.87]	[.23]	[.03]	[.03]	[.04]	[.04]	[552]	[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; RE*x* = earnings in calendar year 19*x*.

[a] NSW sample as constructed by Lalonde (1986).

[b] The subset of the Lalonde sample for which RE74 is available.

[c] 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.

PSID1–3 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.

	(533)	(620)	(459)	(551)	(551)	(671)	(746)	(574)	(662)	(666)	(671)	(672)	(574)	(654)	(65
CPS-3	−1,008	−1	−1,204	−263	−234	−635	375	−91	844	875	−635	1,270	−91	1,326	1,3
	(539)	(681)	(532)	(677)	(675)	(657)	(821)	(641)	(808)	(810)	(657)	(798)	(641)	(796)	(79

NOTES: Panel A replicates the sample of Lalonde (1986, table 5). The estimates for columns (1)–(4) for NSW, PSID1–3, 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 t
obtained did not contain all of the covariates used in column (10) of Lalonde's Table 5.

[a] Estimated effect of training on RE78. Standard errors are in parentheses. The estimates are in 1982 dollars.

[b] The estimates based on the NSW control groups are unbiased estimates of the treatment impacts for the original sample ($886) and for the RE74 sample ($1,794).

[c] 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).

[d] Regresses RE78 on a treatment indicator and RE75.

[e] The same as (d), but controls for the additional variables listed under (c).

[f] Controls for all pretreatment covariates.

Covariate imbalance

- 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
- DW note that the two samples have *severe* imbalance on *observables* – a huge number of non-experimental controls have propensity scores almost exactly equal to 0
- Their analysis will "trim" (which will ultimately have implications for interpretation)

Figure 1. Histogram of the Estimated Propensity Score for NSW Treated Units and PSID Comparison Units. The 1,333 PSID units wh estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin contains 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 treated units with an estimated propensity score greater than .8 and only 7 comparison units.

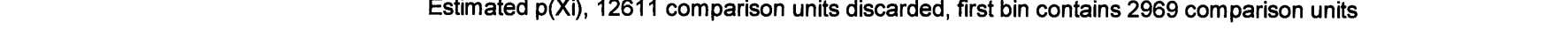


Figure 2. Histogram of the Estimated Propensity Score for NSW Treated Units and CPS Comparison Units. The 12,611 CPS units wh estimated propensity score is less than the minimum estimated propensity score for the treatment group are discarded. The first bin cont 2,969 CPS units. There is minimal overlap between the two groups, but the overlap is greater than in Figure 1; only one bin (.45–.5) contains comparison units, and there are 35 treated and 7 comparison units with an estimated propensity score greater than .8.

	(670)	(658)	(847)	(1,461)	(1,346)		(1,205)	(753)
CPS-3[g]	−635	1,326	556	1,252	2,219	514	587	662
	(657)	(798)	(951)	(1,617)	(2,082)		(1,496)	(776)

[a] Least squares regression: RE78 on a constant, a treatment indicator, age, age^2, education, no degree, black, Hispanic, RE74, RE75.

[b] 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).

[c] 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 propensi
score is greater than the minimum, and less than the maximum, estimated propensity score for the treatment group.

[d] 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:

[e] PSID-1: Prob (T_i = 1) = F(age, age^2, education, education^2, married, no degree, black, Hispanic, RE74, RE75, RE74^2, RE75^2, u74*black).

[f] PSID-2 and PSID-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE74^2, RE75, RE75^2, u74, u75).

[g] CPS-1, CPS-2, and CPS-3: Prob (T_i = 1) = F(age, age^2, education, education^2, no degree, married, black, Hispanic, RE74, RE75, u74, u75, education*RE74, age^3).

		[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.

Replies by econometricians to DW

- Heckman, Smith and Todd concluded from their own work that in order for matching estimators to have low bias, you need the following:
  1. A rich set of variables related to program participation and predictive of Y^0 labor market outcomes,
  2. Nonexperimental comparison group be drawn from the same local labor markets as the participants and
  3. Dependent variable (e.g., 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)
- DW also note the importance of conditioning on pre-treatment lagged outcomes (e.g., real earnings in $t-1$, $t-2$, etc.) as well as *trimming*

Smith and Todd, diff-in-diff, doubly robust

- Difference-in-differences with propensity scores tended to work well in Smith and Todd (2005) though the effect sizes are much larger
- In my Causal Inference II workshop, we use Sant'anna And Zhao's double robust DiD and get nearly the exact same parameter estimate as the experimental finding

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
   - $\rightarrow$ Schooling $\rightarrow$ 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

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

  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?
  - $\rightarrow$ By itself, it's meaningless. It's a reference point between matching solutions.
  - $\rightarrow$ 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.
  - $\rightarrow$ 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.

Roadmap

Comments

- Selection on observables are important and when running regressions with controls, you are in fact doing it
- Conditional independence requires that you *know* and *include* all confounders to adjust comparisons when estimating treatment effects
- Without a prior behavioral model guiding you, it's very hard to defend conditional independence (borderline disingenuous)
- If you are unwilling to use DAGs, you may want to ask yourself why you are comfortable running regressions with covariates?

When not to use selection on observables

- Conditions for selection on observables are strong and subtle
  1. $(Y^1, Y^0) \perp\!\!\!\perp D|X$ (conditional independence)
  2. $0 < Pr(D = 1|X) < 1$ with probability one (common support)
- What exactly does the first thing mean?
  - → Easy to explain part: all the confounders are known and in your dataset
  - → Not as easy to explain part: once you condition on those things, your customers or people were behaving *randomly*
- A lot of matching and weighting focused on biases created by dimensionality problems, but fixing those does not fix first point (and King and Nielsen are only focused on second)

Rationality issues

- Roy models basically suggest people usually do things because it benefits them – called "selection on gains"
  - → If I do something, it's because $Y^1 - Y^0 > 0$
  - → If I don't, it's because $Y^1 - Y^0 < 0$
- We tend to think of this as nearly identical to rationality or intentional behavior, and when we use selection on observables methods we assume that that rationality could be absorbed by observables
- Very mysterious idea and Smith and Sweetman (2015) suggest some kinds of behavior may not *ever* satisfy this condition

Comments

- One diagnostic feature of the matching, weighting and imputation methods over OLS is the steps involved to evaluate common support
- Unnecessary though – you can incorporate the propensity score into regressions as weights
- Nevertheless as we saw with DW, simply trimming can address some of the problems with overlap and propensity score makes this easier by collapsing K strata into a single scalar
- Histograms help to diagnoses these problems

Comments

- Don’t hyper-critical or naive – CIA may or may not hold in your data. You can have too strong of priors in either direction
- All that selection on observables does is create “look-a-likes” on *observables*, but if you left out a critical confounder, you’ve not fixed anything
- Adjusting for covariates may still be valuable even if you are worried about confounders as it can at least you know the differences are due to these known confounders

Brief conclusion

“All models are wrong but some are useful” – George Box

- Keep in mind – you need to know your data, the area you’re in, the people you’re studying.
- Data, credentials and classwork are not a substitute for common sense and thoughtfulness.
- Sometimes CIA isn’t insane and sometimes it may be and there is no rule I can give you

Temporary page!

LaTeX was unable to guess the total number of pages correctly.
was some unprocessed data that should have been added to
page this extra page has been added to receive it.
If you rerun the document (without altering it) this surplus pag
away, because LaTeX now knows how many pages to expect fo
document.