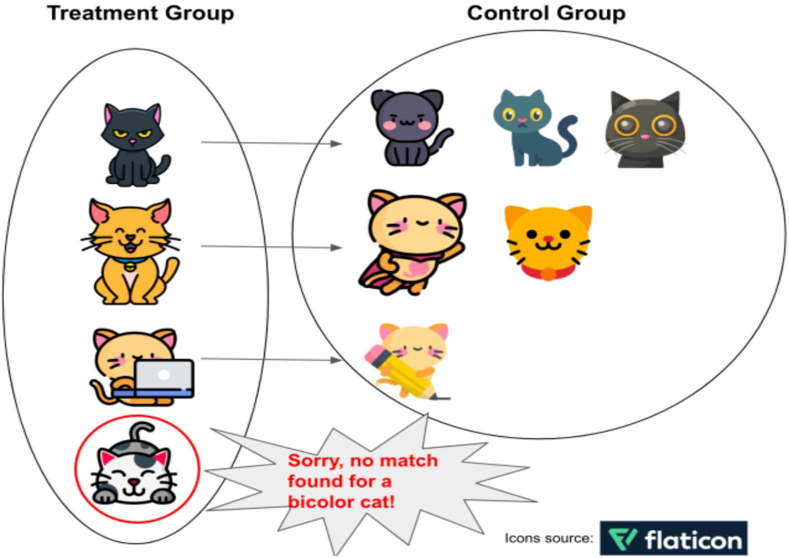
Causal Inference: Matching

POLI 803 Research Methods in PS

Howard Liu

Week 10, 2024

The Idea of Matching



Icons source: flaticon

Estimators

- Now we will explore estimators that for lack of a better word “use covariates” to estimate aggregate causal parameters
- A few topics: subclassification, exact matching, inexact/approximate matching

Roadmap

1. Subclassification

2. Exact Matching

3. Inexact Matching
    - Nearest Neighbor Covariate Matching
    - Propensity scores
        - p-score matching
        - IPW
        - Coarsened Exact Matching (CEM)
        - Coarsened Exact Matching (CEM)

Subclassification method

- Subclassification: an early effort to tackle covariates, but which ultimately cannot handle large features due to common support issues

Subclassification method

- Subclassification: an early effort to tackle covariates, but which ultimately cannot handle large features due to common support issues → no one uses it. shown for pedagogical purpose

Subclassification method

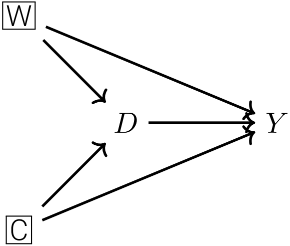
- Subclassification: an early effort to tackle covariates, but which ultimately cannot handle large features due to common support issues → no one uses it. shown for pedagogical purpose
- The Titanic example in the book:

Subclassification method

- Subclassification: an early effort to tackle covariates, but which ultimately cannot handle large features due to common support issues → no one uses it. shown for pedagogical purpose
- The Titanic example in the book:
- Titanic sank
- 2200 on board, but only 700 survived
- **Women** and **children** first was a maritime rule to ration lifeboats, but there were different cabins (**1st class**, 2nd class, etc.) on different levels with different proximity to boats
- Q: What was the causal effect of 1st class on survival after adjusting for W and C?
- Can you draw a DAG (Directed Ayclic graph)?

Exercise: Titanic DAG

Figure: Women W and children C first maritime rule is a confounder for estimating first class D effect on surviving Y



Backdoor criterion can be satisfied by blocking on W and C. These are our known confounders. Now we just need data to see if it's quantified.

Titanic exercise

1. **Stratify the confounders**: Our age and sex variables are both binary, so we can only create four strata: male children, female children, male adults, female adults → to ”control” the covariates

Titanic exercise

1. **Stratify the confounders**: Our age and sex variables are both binary, so we can only create four strata: male children, female children, male adults, female adults → to ”control” the covariates
2. **Calculate differences within strata**: Calculate average survival rates for each group within each of the four strata and difference within strata

Titanic exercise

1. **Stratify the confounders**: Our age and sex variables are both binary, so we can only create four strata: male children, female children, male adults, female adults → to ”control” the covariates
2. **Calculate differences within strata**: Calculate average survival rates for each group within each of the four strata and difference within strata
3. **Calculate probability weights**: Count the number of people in each strata and divide by the total number of souls aboard (crew and passengers)

Titanic exercise

1. **Stratify the confounders**: Our age and sex variables are both binary, so we can only create four strata: male children, female children, male adults, female adults $\rightarrow$ to ”control” the covariates
2. **Calculate differences within strata**: Calculate average survival rates for each group within each of the four strata and difference within strata
3. **Calculate probability weights**: Count the number of people in each strata and divide by the total number of souls aboard (crew and passengers)
4. **Aggregate differences across strata using weights**: Estimate the ATE by aggregating the difference in survival rates over the four strata with each strata-specific difference weighted by that strata’s weight

Table 1: Stratified sample

Table: Counts and Titanic survival rates by strata and first class status.

	First class		**All other classes**		
Strata	Obs	Mean	Obs	Mean	Total
Male adult	175	0.326	1,492	0.188	1,667
Female adult	144	0.972	281	0.626	425
Male child	5	1	59	0.407	64
Female child	1	1	44	0.613	45
Total observations	325		1,876		2,201

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification		**Stratification weighted estimates**	
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates?

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification		**Stratification weighted estimates**	
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification		**Stratification weighted estimates**	
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137
3. Prob weights for ATE?

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification		**Stratification weighted estimates**	
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137
3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76
3. Prob weights for ATT?

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76

3. Prob weights for ATT? 175 (treat)/325 (total in treat)= 0.54

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76

3. Prob weights for ATT? 175 (treat)/325 (total in treat)= 0.54

3. Prob weights for ATU?

Table 2: Estimates of aggregate parameters

Strata	Differences in Survival Rates	Weight$_{k,ATE}$	Weight$_{k,ATT}$	Weight$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76

3. Prob weights for ATT? 175 (treat)/325 (total in treat)= 0.54

3. Prob weights for ATU? 1492 (control)/1876 (total in treat)= 0.80

Table 2: Estimates of aggregate parameters

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	0.387	0.02	0.00	0.02

	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	0.20	0.24	0.19

2. Diff in survival rates? 0.326-0.188 = 0.137

3. Prob weights for ATE? 1667 (treat+control)/2201 (all) = 0.76

3. Prob weights for ATT? 175 (treat)/325 (total in treat)= 0.54

3. Prob weights for ATU? 1492 (control)/1876 (total in treat)= 0.80

$$\widehat{\delta}_{ATT} = (0.137 \times 0.54) + (0.346 \times 0.44) + (0.593 \times 0.02) + (0.387 \times 0.00)$$

$$= 0.24 \text{ or 24 percentage points}$$

What if we drop the only female child in the 1st class (due to false data) ?

Table: Differences in survival rates, stratification weights, and estimates of parameters without perfect stratification

Strata	**Differences in Survival Rates**	**Weight**$_{k,ATE}$	**Weight**$_{k,ATT}$	**Weight**$_{k,ATU}$
Male adult	0.137	0.76	0.54	0.80
Female adult	0.346	0.19	0.44	0.15
Male child	0.593	0.03	0.02	0.03
Female child	n/a	n/a	n/a	0.02
	No stratification	**Stratification weighted estimates**		
	$\widehat{\textbf{SDO}}$	$\widehat{\textbf{ATE}}$	$\widehat{\textbf{ATT}}$	$\widehat{\textbf{ATU}}$
Estimated coefficient	0.35	n/a	0.24	n/a

Differences in survival rates, stratification weights, and estimated parameters. All coefficients should be multiplied by 100 to get a percentage point change in survival rate as a result of having a first class cabin. Note that the SDO is a simple difference in mean outcomes and therefore *not* a weighted average over the strata differences. But the estimated ATE, ATT and ATU parameters are weighted averages in difference in means using corresponding stratification weights.

Lack of Common Support (Empty cells)

- Stratification requires having units in both groups for every value of X to get ATE
- If you want the ATT, you have to have units in the control group for every treated group based on its value of X (female children weren't treated after n/a, so didn't matter)

Lack of Common Support (Empty cells)

- Stratification requires having units in both groups for every value of X to get ATE
- If you want the ATT, you have to have units in the control group for every treated group based on its value of X (female children weren't treated after n/a, so didn't matter)
- If you want the ATU, you have to have units in the treatment group for every treated group based on its value of X (female children weren't treated, so it "did" matter)

Lack of Common Support (Empty cells)

- Stratification requires having units in both groups for every value of X to get ATE
- If you want the ATT, you have to have units in the control group for every treated group based on its value of X (female children weren't treated after n/a, so didn't matter)
- If you want the ATU, you have to have units in the treatment group for every treated group based on its value of X (female children weren't treated, so it "did" matter)
- This has a technical word we are going to learn more about called a "**lack of common support**" (missings obs)

No overlap (common support violation)

No support is like an incomplete bridge which stops you from even being able to cross the moat even though the troops exist (i.e., unconfoundedness)

Curse of Dimensionality

- Dimensionality (k) usually means number of covariates. Imagine the problem here?

Curse of Dimensionality

- Dimensionality (k) usually means number of covariates. Imagine the problem here?
- Stratification methods break down in finite samples because as increase the number of covariates, the ”dimension” grows even faster
- 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 curse of dimensionality is based on the slices of all interactions of the covariates, not just the covariates, and that explodes fast $\rightarrow$ we need a better method to create strata

Roadmap

1. Subclassification

2. Exact Matching

3. Inexact Matching
    - Nearest Neighbor Covariate Matching
    - Propensity scores
        - p-score matching
        - IPW
        - Coarsened Exact Matching (CEM)
        - Coarsened Exact Matching (CEM)

Exact matching

Exact Matching

- Subclassification: uses the difference between treatment and control group units and achieves covariate balance by using the k probability weights to weight the averages

Exact Matching

- Subclassification: uses the difference between treatment and control group units and achieves covariate balance by using the k probability weights to weight the averages
  $\rightarrow$ failed quickly when k goes large and many missing matches (common support failed)

The Training example: Age and Earnings

- How do we create matches/counterfactuals with the treated group?

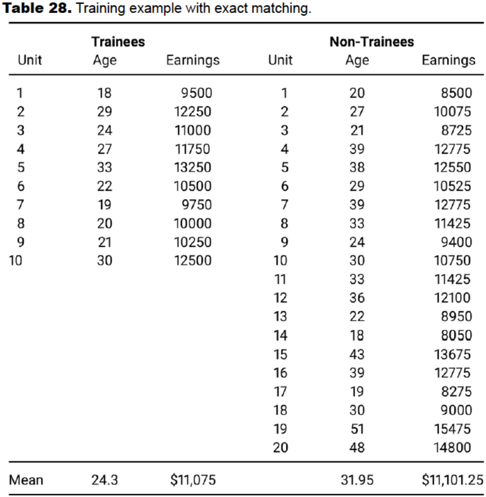


Table 28. Training example with exact matching.

Trainees			Non-Trainees		
Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500
2	29	12250	2	27	10075
3	24	11000	3	21	8725
4	27	11750	4	39	12775
5	33	13250	5	38	12550
6	22	10500	6	29	10525
7	19	9750	7	39	12775
8	20	10000	8	33	11425
9	21	10250	9	24	9400
10	30	12500	10	30	10750
			11	33	11425
			12	36	12100
			13	22	8950
			14	18	8050
			15	43	13675
			16	39	12775
			17	19	8275
			18	30	9000
			19	51	15475
			20	48	14800
Mean	24.3	$11,075		31.95	$11,101.25

Exact matching

- Exact matching finds a person in the control group whose value of X_j is *exactly* equal to each person in the treatment group i
- Will not work if the conditioning set includes a continuous variable
- Will also not work if K gets large (curse of dimensionality we discuss later)

The Training Example

- Take average with multiple exact matches ($\frac{1}{M}$)

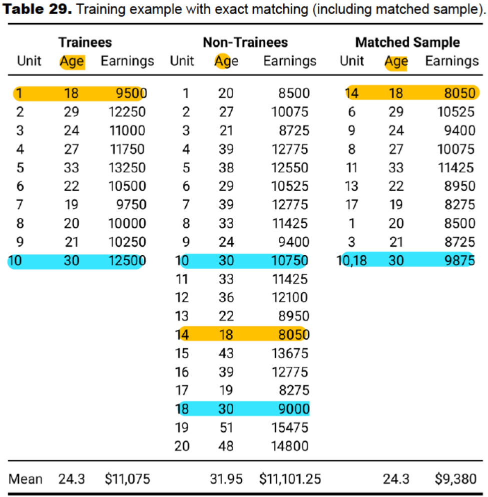


Table 29. Training example with exact matching (including matched sample).

Trainees			Non-Trainees			Matched Sample		
Unit	Age	Earnings	Unit	Age	Earnings	Unit	Age	Earnings
1	18	9500	1	20	8500	14	18	8050
2	29	12250	2	27	10075	6	29	10525
3	24	11000	3	21	8725	9	24	9400
4	27	11750	4	39	12775	8	27	10075
5	33	13250	5	38	12550	11	33	11425
6	22	10500	6	29	10525	13	22	8950
7	19	9750	7	39	12775	17	19	8275
8	20	10000	8	33	11425	1	20	8500
9	21	10250	9	24	9400	3	21	8725
10	30	12500	10	30	10750	10,18	30	9875
			11	33	11425			
			12	36	12100			
			13	22	8950			
			14	18	8050			
			15	43	13675			
			16	39	12775			
			17	19	8275			
			18	30	9000			
			19	51	15475			
			20	48	14800			
Mean	24.3	$11,075		31.95	$11,101.25		24.3	$9,380

ATT estimator: One match

We will focus on the ATT for the rest of today and the equation is:

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

where $Y_{j(i)}$ is the j^{th} unit matched to the i^{th} unit based on the j^{th} being "exactly equal to" the i^{th} unit with respect to the X conditioning set

Multiple matches

Multiple matches: What if I find two or more M units with the identical X value? Then what?

Multiple matches

Multiple matches: What if I find two or more M units with the identical X value? Then what?

$$\widehat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D_i=1} \bigg(Y_i - \bigg[\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(1)} \bigg] \bigg) \tag{2}$$

Notice that we are only dealing with Y_i^0 (control) by matching; The Y_i^1 (treatment) is fine as is.

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, **find** all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, **find** all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.
2. For each unit i, **replace its missing potential outcome**, Y_i^0, with the matched j units' realized outcomes, $\frac{1}{M}\sum Y_{j(i)}$, from Step 1. Do this for all i units in the treatment group.

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, **find** all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.
2. For each unit i, **replace its missing potential outcome**, Y_i^0, with the matched j units' realized outcomes, $\frac{1}{M}\sum Y_{j(i)}$, from Step 1. Do this for all i units in the treatment group.
3. For each unit i, **calculate the difference** between realized earnings and matched earnings, $\widehat{\delta}_i = Y_i - \frac{1}{M}\sum Y_{j(i)}$.

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, **find** all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.
2. For each unit i, **replace its missing potential outcome**, Y_i^0, with the matched j units' realized outcomes, $\frac{1}{M}\sum Y_{j(i)}$, from Step 1. Do this for all i units in the treatment group.
3. For each unit i, **calculate the difference** between realized earnings and matched earnings, $\widehat{\delta}_i = Y_i - \frac{1}{M}\sum Y_{j(i)}$.
4. Finally, estimate the sample ATT by averaging over all i differences in earnings from Step 3 as $\frac{1}{N_T}\sum \widehat{\delta}_i$, where N_T is the number of treatment units.
5. Easy peasy!

Matching algorithm

1. For each unit i in the treatment group with known and quantified confounder $X = x_i$, **find** all units j in the donor pool for whom $x_i = x_j$. These j units are our M matches and M can be one or it can be greater than one if you want it to be.
2. For each unit i, **replace its missing potential outcome**, Y_i^0, with the matched j units' realized outcomes, $\frac{1}{M}\sum Y_{j(i)}$, from Step 1. Do this for all i units in the treatment group.
3. For each unit i, **calculate the difference** between realized earnings and matched earnings, $\widehat{\delta}_i = Y_i - \frac{1}{M}\sum Y_{j(i)}$.
4. Finally, estimate the sample ATT by averaging over all i differences in earnings from Step 3 as $\frac{1}{N_T}\sum \widehat{\delta}_i$, where N_T is the number of treatment units.
5. Easy peasy!
6. What problem we would encounter here?

Roadmap

1. Subclassification

2. Exact Matching

3. Inexact Matching

  - Nearest Neighbor Covariate Matching
  - Propensity scores
    - p-score matching
    - IPW
    - Coarsened Exact Matching (CEM)
    - Coarsened Exact Matching (CEM)

Inexact matching

Nearest Neighbor Covariate Matching (NN Matching)

- What if you couldn't find another unit with that exact same value of X?

Nearest Neighbor Covariate Matching (NN Matching)

- What if you couldn’t find another unit with that exact same value of X? $\rightarrow$ Then you do approximate matching

Nearest Neighbor Covariate Matching (NN Matching)

- What if you couldn't find another unit with that exact same value of X? → Then you do approximate matching
- Estimate $\widehat{\delta}_{ATT}$ by **imputing** the missing potential outcome of each treatment unit i using the observed outcome from that outcome's "nearest" neighbor j in the control set using X for the matching

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

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:

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

- 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 $M = 2$)

Example: 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 $\widehat{\delta}_{ATT} = \dfrac{1}{N_T} \displaystyle\sum_{D_i=1} (Y_i - Y_{j(i)})$?

Example: 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

ATT?

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

ATT?

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

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

Measuring the matching discrepancy

- What does it mean to be close when I am working with a large number of covariates $k > 1$?
- Need a way of measuring a match in terms of how “close” each unit’s X_i value was to the matched X_j

Measuring the matching discrepancy

- What does it mean to be close when I am working with a large number of covariates $k > 1$?
- Need a way of measuring a match in terms of how “close” each unit’s X_i value was to the matched X_j
  $\rightarrow$ Let’s do that and use the square root of the sum of all squared differences in each unit’s $X_i - X_{j(i)}$ as a measure of how bad the match is
  $\rightarrow$ This is called the Euclidean distance

Euclidean distance

Definition: Euclidean distance

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

Minimizing the Euclidean distance

- Abadie and Imbens (2006) show that there exists a unique solution to the matching problem that minimizes a given distance metric
- `Matching` in R and `teffects` in Stata (not sure in python)
- But the idea here is that any other match will always have a higher Euclidean distance

Other distance metrics

- Our example treated a one unit difference in age and one unit difference in GPA as the same, but those scales are different and matter a lot
- The Euclidean distance is not invariant to changes in the scale of the X's.
- Alternative distance metrics that *are* invariant to changes in scale are more commonly used
- Normalized Euclidean distance and Mahalanobis distance both try to **normalize** it so that scale doesn't matter

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:

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

- 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.

Matching and the Curse of Dimensionality

Matching and the Curse of Dimensionality

- Common support and the dimensality curse: The larger the dimensions of the conditioning set, the less likely common support holds, and you can’t not do it because you need these covariate dimensions to satisfy weak unconfoundedness!
- This problem is caused by the finite dataset, and it introduces a particular type of selection bias
- Curses are only overcome with new spells!

Matching and the Curse of Dimensionality

- Common support and the dimensality curse: The larger the dimensions of the conditioning set, the less likely common support holds, and you can’t not do it because you need these covariate dimensions to satisfy weak unconfoundedness!
- This problem is caused by the finite dataset, and it introduces a particular type of selection bias
- Curses are only overcome with new spells! → Abadie and Imbens (2011) derived a way to reduce the bias (bias adjustment or bias correction) → in the `matchIt` package

Curse of dimensionality, bias and heterogeneous treatment effects

- Recall the problem of many covariates for exact matching – the curse of dimensionality makes matching on K covariates implausible as the dimensions grow exponentially with K
- This is problem because recall there are two assumptions needed to match
    1. Unconfoundedness: this gives you the right to match
    2. Common support: this gives you the ability to match
- Without both, then depending on the amount of hetergeneity in the treatment effects, matching will be biased

Curse of dimensionality, bias and heterogeneous treatment effects

- Recall the problem of many covariates for exact matching – the curse of dimensionality makes matching on K covariates implausible as the dimensions grow exponentially with K
- This is problem because recall there are two assumptions needed to match
  1. Unconfoundedness: this gives you the right to match
  2. Common support: this gives you the ability to match
- Without both, then depending on the amount of hetergeneity in the treatment effects, matching will be biased

  $\rightarrow$ We introduce a different technique to address this curse of dimension

Propensity score as dimension reduction

- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method
- Propensity score theorum: They show that if treatment is independent of potential outcomes conditional on K covariates, then it will be independent of potential outcomes conditional on propensity score

Propensity score as dimension reduction

- Rubin (1977) and Rosenbaum and Rubin (1983) developed the propensity score method
- Propensity score theorum: They show that if treatment is independent of potential outcomes conditional on K covariates, then it will be independent of potential outcomes conditional on propensity score
- Main value of the propensity score is dimension reduction to reduce K covariates into a **single scalar** without loss of information
- Variety of ways to incorporate the propensity score – stratification, weighting and matching. We focus on propensity score matching here.

Caveat: **Propensity score** matching

- King and Nielson (2019) “Why propensity scores should not be used for matching”
- “The more balanced the data (or the more balanced it comes by trimming some of the observations through matching), the more likely propensity score matching will degrade inferences.”

Caveat: **Propensity score** matching

- King and Nielson (2019) "Why propensity scores should not be used for matching"
- "The more balanced the data (or the more balanced it comes by trimming some of the observations through matching), the more likely propensity score matching will degrade inferences."
- → caution against assuming that more balance always improves causal inference
- Reduced sample size
- Bias from exclusion of extreme values
- Overfitting to the matched sample
- The author thinks its fine. I see less work on p-score matching these days

Propensity score matching: three core steps

1. Takes necessary covariates

Propensity score matching: three core steps

1. Takes necessary covariates
2. Estimates a maximum likelihood model of the conditional probability of treatment (usually a logit or probit so as to ensure that the fitted values are bounded between 0 and 1),

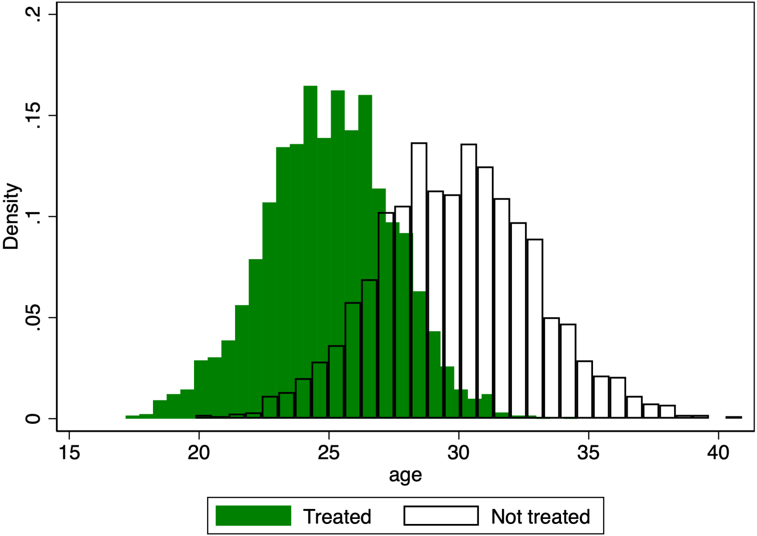
Propensity score matching: three core steps

1. Takes necessary covariates
2. Estimates a maximum likelihood model of the conditional probability of treatment (usually a logit or probit so as to ensure that the fitted values are bounded between 0 and 1),
3. Uses the predicted values from that estimation to collapse those covariates into a single scalar called the propensity score. All comparisons between the treatment and control group are then based on that value (propensity).

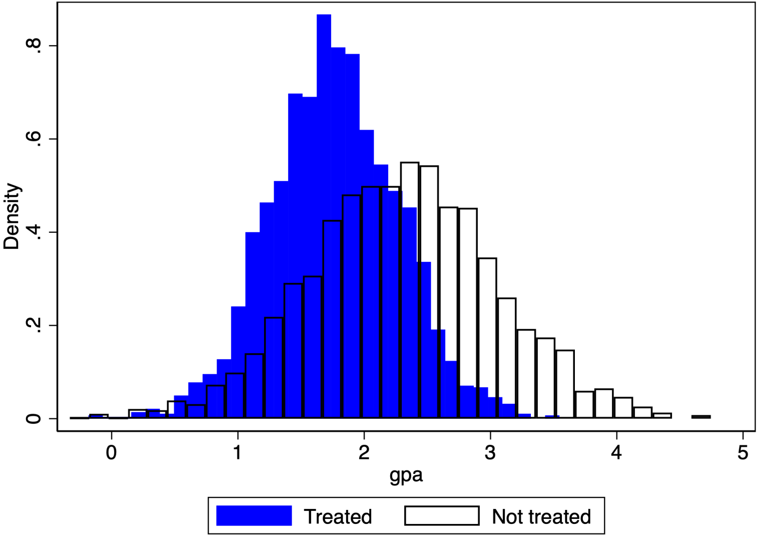
Investigating overlap

- Once you obtain the propensity score, you can use it for estimation, but you can also use it for evaluating covariate balance
- It's an easy way to do it with histograms, and since the propensity score theorem holds for the dimensions of X, there's no loss of generality in investigating overlap that way versus one by one
- Remember: you need common support, not on X individual covariates alone, but within the K dimensions, so investigating with the propensity score is easier to do

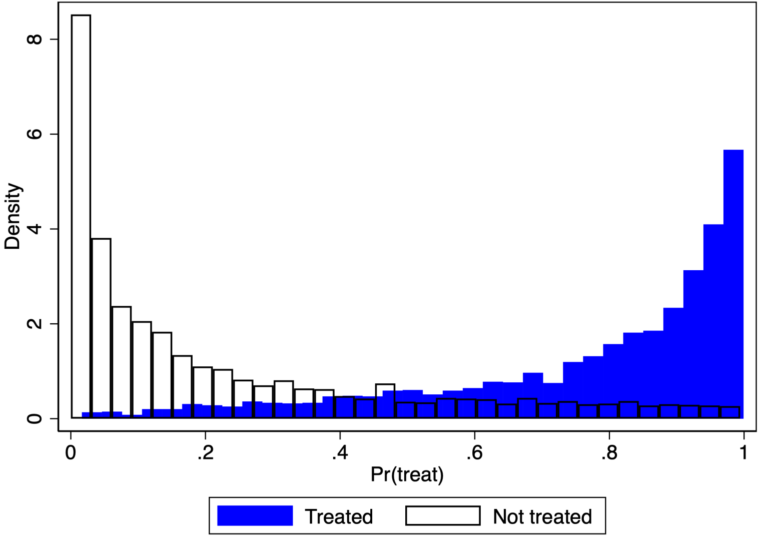
Covariate 1 histograms: Age



Covariate 2 histograms: GPA



Summarizing both with propensity score histogram



Navigating the vastness of estimation

- Estimators abound and can be a little bewildering so to summarize them:
  1. Match units from one group to another **using the propensity score** (with various rules for finding how close to be)
  2. Weighting by the inverse propensity score
- Variety of techniques to derive standard errors from parametric methods to bootstrapping
- Can even introduce “doubly robust” methods to deal with matching bias like we did with nearest neighbor bias correction
- But all these estimators assume unconfoundedness, so really it’s all about addressing the lack of overlap; that’s the only bias that exists when you have unconfoundedness

Step 1: Pick your parameter ATE vs ATT vs ATU

- Which population are you studying? Only those who were discriminated against? That’s the ATT

Step 1: Pick your parameter ATE vs ATT vs ATU

- Which population are you studying? Only those who were discriminated against? That’s the ATT
- Do you want to imagine “what if blacks and whites were both discriminated against?” That’s the ATE

Step 1: Pick your parameter ATE vs ATT vs ATU

- Which population are you studying? Only those who were discriminated against? That's the ATT
- Do you want to imagine "what if blacks and whites were both discriminated against?" That's the ATE
- The more you can focus on one particular causal parameter, the easier and more justified it gets as it weakens both assumptions (unconfoundedness and common support)

Step 2: Estimate the propensity score

- Estimate the conditional probability of treatment using probit or logit model (or ML)
$$Pr(D_i = 1|X_i) = F(\beta X_i)$$
- Note: don't use OLS because while it will get the mean right, it will not get correct values in the tails because of its linear projections
- OLS will give propensity scores outside the [0,1] bounds and probabilities cannot be negative or greater then one

Step 2: Estimate the propensity score

- Use the estimated coefficients to predict 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

Step 2: Estimate the propensity score

- Think of the propensity score as a frequentist concept of probability
- “If I drew someone from the sample with these characteristics, then how many of those are in the treatment group divided by the total with those characteristics”
- Or for each dimension of X, a ratio of $\frac{N_T}{(N_T+N_C)}$

Step 3a: Propensity Score Estimation with matching

- Most common method is to use **matching** (propensity score matching or matching on propensity scores)
- Matching finds a unit in the comparison group with a similar $\widehat{\rho}_i(X)$ to service as counterfactual for the unit

Step 3a: Propensity Score Estimation with matching

- Most common method is to use **matching** (propensity score matching or matching on propensity scores)
- Matching finds a unit in the comparison group with a similar $\widehat{\rho}_i(X)$ to service as counterfactual for the unit
- For the ATE, you'll need matches on both side; for the ATT, you'll need matches for the treatment group among controls

Step 3a: Propensity Score Estimation with matching

- Most common method is to use **matching** (propensity score matching or matching on propensity scores)
- Matching finds a unit in the comparison group with a similar $\widehat{\rho}_i(X)$ to service as counterfactual for the unit
- For the ATE, you'll need matches on both side; for the ATT, you'll need matches for the treatment group among controls
- Lack of overlap creates issues for matching which we'll note later

Step 3b: Propensity Score Estimation with stratification

- Rare to see this done anymore, though it was one of the methods that Dehejia and Wahba (2002) tried
- Stratification is a kind of weighting method similar to Cochran's subclassification method where weights are group shares **within certain ranges of the propensity score**

Step 3c: Early weighting methods

- Called Weighting on the propensity score.

Step 3c: Early weighting methods

- Called Weighting on the propensity score.
- Most common is the inverse probability weighting (IPW)

Step 3c: Estimation with inverse probability weighting

- IPW uses the estimated propensity score to reweight the outcomes for which there are several historical methods for doing so
- IPW is non-parametric – you are just taking averages and multiplying by the inverse of the propensity score weights depending on which parameter you want to estimate
- There are fewer implementation choices than in matching (i.e., no choice over distance, number of neighbors)
- There are bias adjustment methods called double robust where you combine imputing counterfactuals with weighting by the propensity score

Step 3d: Coarsened Exact Matching

- Also called weighting on the propensity score
- Iacus et al. (2012) introducted it
- Very simple idea: it’s possible to do exact matching once we **coarsen the data enough.**
- E.g., 0- to 10-year-olds, 11- to 20-year olds, then oftentimes we can find exact matches

Step 3d: Coarsened Exact Matching

- Also called weighting on the propensity score
- Iacus et al. (2012) introducted it
- Very simple idea: it's possible to do exact matching once we **coarsen the data enough.**
- E.g., 0- to 10-year-olds, 11- to 20-year olds, then oftentimes we can find exact matches

Coarsened Exact Matching: Steps

1. We begin with covariates X and make a copy called $X*$.

Coarsened Exact Matching: Steps

1. We begin with covariates X and make a copy called $X*$.
2. Next we coarsen $X*$ according to user-defined cutpoints or **CEM's automatic binning algorithm.** For instance, schooling becomes less than high school, high school only, some college, college graduate, post college

Coarsened Exact Matching: Steps

1. We begin with covariates X and make a copy called $X*$.
2. Next we coarsen $X*$ according to user-defined cutpoints or **CEM's automatic binning algorithm.** For instance, schooling becomes less than high school, high school only, some college, college graduate, post college
3. Then we create one stratum per unique observation of $X*$ and place each observation in a stratum. Assign these strata to the original and uncoarsened data, X, and drop any observation whose stratum doesn't contain at least one treated and control unit.
4. Then add weights for stratum size and analyze without matching (similar to what we did in the subclassfication example)

Step 4: Standard Errors

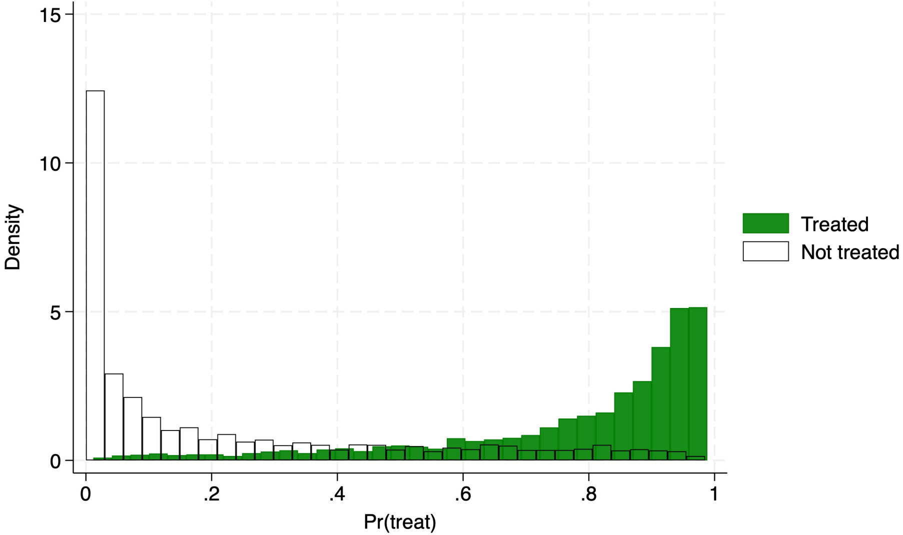
Standard errors can be constructed a few different ways:

- 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)

Reminder: Check for common support assumption using histograms

- Assessing whether there are units it both groups for whichever parameter you're focused on is simple with propensity score as shown earlier using histograms of the propensity score for treated and control → little overlaps mean bad matching
- Crump, et al. (2009) suggest keeping propensity scores within the interval [0.1,0.9] ("trimming") but any trimming will drop units and dropping units means moving away from the parameter
- Let's look at a picture again just to remind ourselves

Assessing overlap


Density
Pr(treat)
Treated
Not treated