ECON 672

Week 4: Matching and Propensity Scores

Samuel Rowe, PhD 12/19/2022

Overview

- Takeaway
- Conditional Independence Assumption
- Subclassification
- Matching Identification Strategies
  - Exact Matching
  - Approximate Matching
    - Distance Matching
    - Propensity Score Matching
    - Coarsen Exact Matching

Overview

- There are there generalized conditioning (matching) strategies that we will cover
- Subclassification
  - The least likely to be used
- Exact Matching
  - Works well when you don’t have a lot of confounders and lots of data
- Approximate Matching
  - This is the most common way to match, especially using propensity scores

The Takeaway

General Conditioning (Matching) Strategies

- Strengths
  - Selection on observables (observable confounders)
  - Very common strategy in federal evaluation
- Weaknesses
  - No selection on unobservables (unobservable confounders)
  - Without an RCT to benchmark, our matching strategy is less credible identification strategy

The Takeaway

- Assumptions
  - Conditional Independence Assumption
  - Common Support Assumption
- Testable Assumptions
  - Common Support Assumption

Takeaway


OUR MATCHING METHOD CONTROLS FOR CONFOUNDERS
AND UNOBSERVED CONFOUNDERS
AND UNOBSERVED CONFOUNDERS, RIGHT?
imgflip.com

Conditional Independence Assumption

Conditional Independence Assumption

- The main assumption in the conditioning (matching) strategies is the conditional independence assumption
  - Treatment is independent potential outcomes when conditioned on a vector of covariates
  - $(Y^1, Y^0) \perp D \,|\, X$

Conditional Independence Assumption

- $(Y^1, Y^0) \perp D \,|\, X$
- This assumption means that the expected potential outcomes Y1 and Y0 are equal between treatment and control groups
  - $E[Y^1 \,|\, D = 1, X] = E[Y^1 \,|\, D = 0, X]$
  - $E[Y^0 \,|\, D = 1, X] = E[Y^0 \,|\, D = 0, X]$

Conditional Independence Assumption

- Two Concepts with Conditional Independence Assumption
- 1)When the Conditional Independence Assumption is satisfied,
  - The backdoor criterion is satisfied and the identification strategy is credible
- 2) Selection on observables
  - Treatment assignment had been conditioned on observed variables

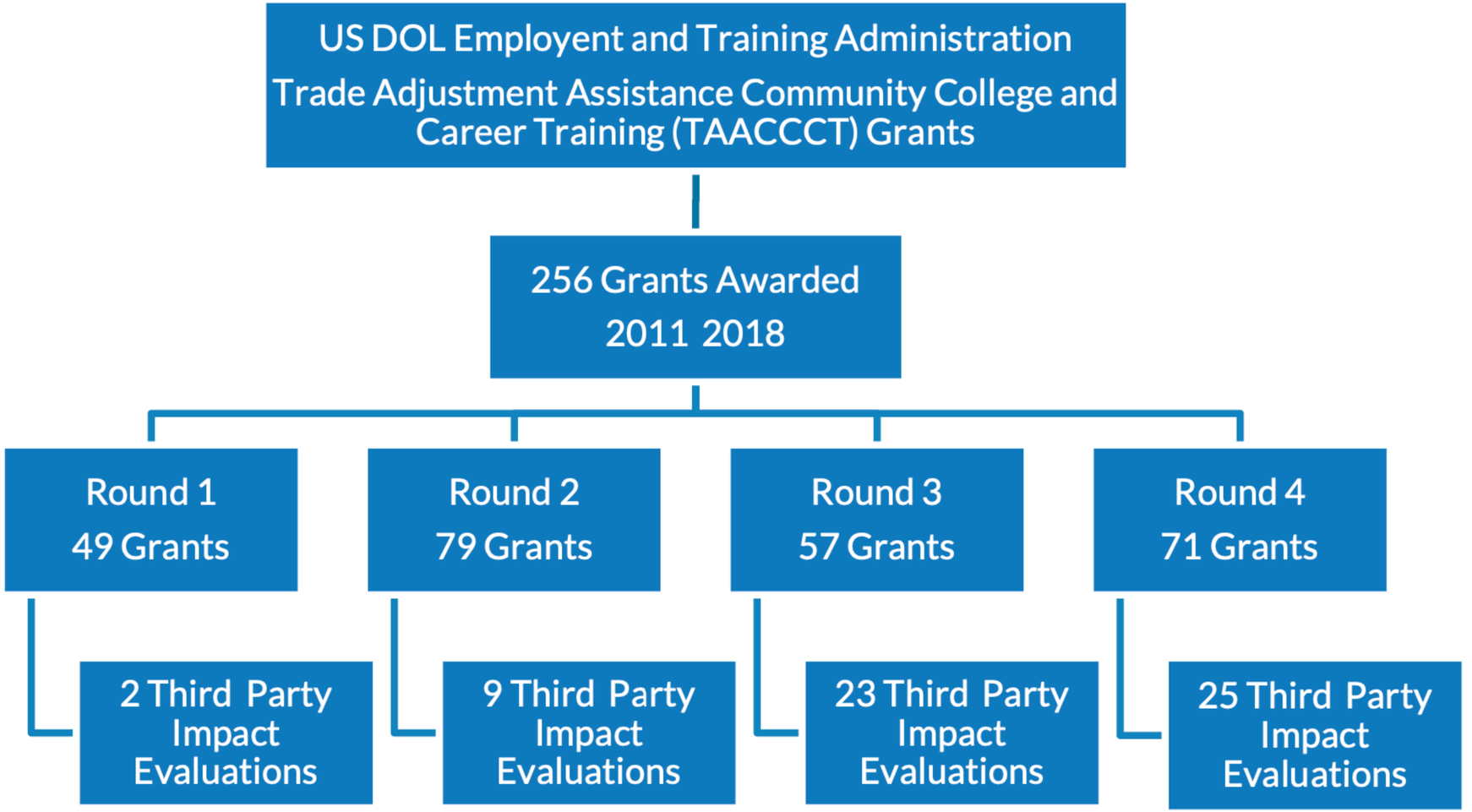
Matching Evaluation Example: TAACCCT

- Trade Adjustment Assistance Community College and Career Training (TAACCCT) Grant Program Evaluation
  - Employment and Training Administration in USDOL award $1.9 billion in grants to 256 grantees between 2011 and 2018
- Impact Evaluation
  - In the 3rd Round of Grants, 57 grantees submitted evaluation plans
  - Only 23 out of the 57 did an evaluation
  - All 23 evaluations used propensity score matching instead of RCT

Matching Evaluation Example: TAACCCT

FIGURE 1.2

Grants Awarded and Third-Party Impact Evaluations Across All Rounds of the TAACCCT Grants



TAACCCT grant project (listed in order of consistently positive impact results), followed by other grant projects listed alphabetically)	Educational outcomes	Employment outcomes
1. Golden Triangle Modern Manufacturing	Positive	Positive
2. IMPACT	Positive	Positive
3. INTERFACE	Positive	Positive
4. Rural Information Technology Alliance	Positive	Positive
5. Advanced Manufacturing, Mechatronics, and Quality Consortium	No impact	Positive
6. BOOST	Mixed	Not studied
7. Bridging the Gap	Mixed	No impact
8. Central Georgia Healthcare Workforce Alliance	Positive	Not studied
9. DC Construction Academy and DC Hospitality Academy	Positive	Not studied
10. Greater Cincinnati Manufacturing Career Accelerator	Mixed	Not studied
11. Health Science Pathways for Academic Career and Transfer Success	Positive	Not studied
12. Linn-Benton iLearn	Positive	No impact
13. Maine is IT!	Mixed	Not studied
14. Mississippi River Transportation, Distribution, and Logistics	Positive[a]	Not studied
15. North Dakota Advanced Manufacturing Skills Training Initiative	Positive	Negative[b]
16. Northeast Resiliency Consortium	Positive	Not studied
17. Orthopedics, Prosthetics, and Pedorthics (HOPE) Careers Consortium	No impact	Not studied
18. PA Manufacturing Workforce Training Center	Not studied	No impact
19. Pathways to Success	Positive	Not studied
20. RevUp	Negative	Not studied
21. Southeastern Economic and Education Leadership Consortium	No impact	No impact
22. Southwest Arkansas Community College Consortium	Positive[a]	Not studied
23. XCEL-IT	Mixed	No impact
Total number of evaluations with positive impacts	13 of 22 studies with educational outcomes	6 of 11 studies with employment outcomes

Subclassification

Subclassification Overview

- Created by Cochran (1968) to assess effects of smoking on lung cancer
  - This was foundation work that setups up other conditioning strategies
- Subclassification is a method that satisfies the backdoor criterion using weighted differences
  - It uses strata-specific weights
- This conditioning strategy achieves distributional balance between treatment and control
  - Uses K number of strata probability weights to weigh the average outcomes

Subclassification Takeaway

- Strength
  - Simple to calculate
- Weakness
  - Problem of dimensionality
    - There is a lack of common support due to small sample sizes
  - Conditional independence assumption might not be satisfied if there are ***unobserved confounders***
- Assumptions
  - Conditional Independence Assumption (not testable)
  - Common Support Assumption (testable)

Subclassification

- Average Treatment Effect Estimator
  - $\hat{\delta}_{ATE} = \int \big(E[Y|D=1,X] - E[Y|D=0,X]\big)dx$
- Assumptions
  - Conditional independence assumption
    - $(Y^1, Y^0) \perp D \,|\, X$
  - Common Support Assumption
    - Probability of being assigned to treatment is between 0 and 1 for each strata
    - $0 < Pr(D=1\,|\,X) < 1$
    - Note that we cannot calculate relevant weights without common support in each strata

Subclassification

- Cochran (1968) wanted to smoking and lung cancer after previous studies showed correlation
    - Prior studies did not establish causation
- People self select into smoking, so smokers and nonsmokers will differ
    - It was possible that smokers had ***confounders*** that created spurious correlations between smoking and lung cancer
- Cochran (1968) uses age-adjusted mortality rates
    - He generated age-specific strata by types of smoking

Subclassification

- A naive comparison shows death rates between smokers and non-smokers to be fairly similar, except for cigar and pipes

Table 5.1: Death rates per 1,000 person-years (Cochran 1968)

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	20.5	14.1	13.5
Cigars/pipes	35.5	20.7	17.4

Subclassification

- Why would cigar/pipe smokers have a higher death rate than cigarette smokers and why would non-smokers and cigarette smokers have similar death rates?
- Do we believe that cigar smoking is more dangerous than cigarette smoking?
  - $E[Y^0 \mid Cigars] = E[Y^0 \mid Cigarettes]$
- Do we believe that cigarette smoking is no more danger than non-smoking?
  - $E[Y^0 \mid Cigarettes] = E[Y^0 \mid NonSmoking]$
- It is unlikely that these independent assumptions hold

Subclassification

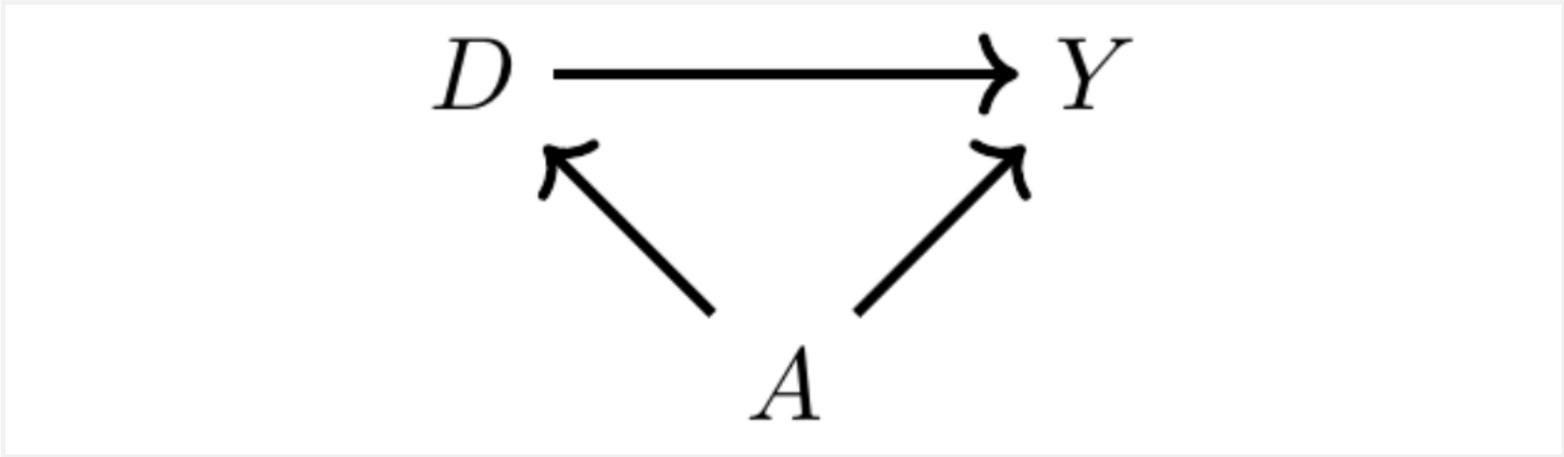
- **Confounder**
  - Age
  - Age is related to types of smoking and death

Table 5.2: Mean ages, years (Cochran 1968).

Smoking group	Canada	British	US
Non-smokers	54.9	49.1	57.0
Cigarettes	50.5	49.8	53.2
Cigars/pipes	65.9	55.7	59.7

Subclassification

- Age is probably not the only confounder, but we will set up our DAG
  - Direct Pathway: $Smoking(D) \rightarrow Deaths(Y)$
  - Indirect Pathway: $Smoking(D) \leftarrow Age(A) \rightarrow Deaths(Y)$



Subclassification

- We need to address the covariate imbalance
  - There is a lack of balance between “treatment” and age
  - Generate age-specific strata

Table 5.3: Subclassification example.

	Death rates	# of Cigarette smokers	# of Pipe or cigar smokers
Age 20–40	20	65	10
Age 41–70	40	25	25
Age $\geq$ 71	60	10	65
Total		100	100

Subclassification

- Our strata-weights are $\frac{N_T}{N}$
- Death Rates of cigarettes without subclassifications for cigarette smokers
  - $20 \times \frac{65_{20-40}}{100} + 40 \times \frac{25_{41-70}}{100} + 60 \times \frac{10_{71+}}{100} = 29 \ per \ 100{,}000$
- Death Rate of cigarettes with subclassifications (age-adjusted)
  - $20 \times \frac{10_{71+}}{100} + 40 \times \frac{25_{41-70}}{100} + 60 \times \frac{65_{20-40}}{100} = 51 \ per \ 100{,}000$

Subclassification

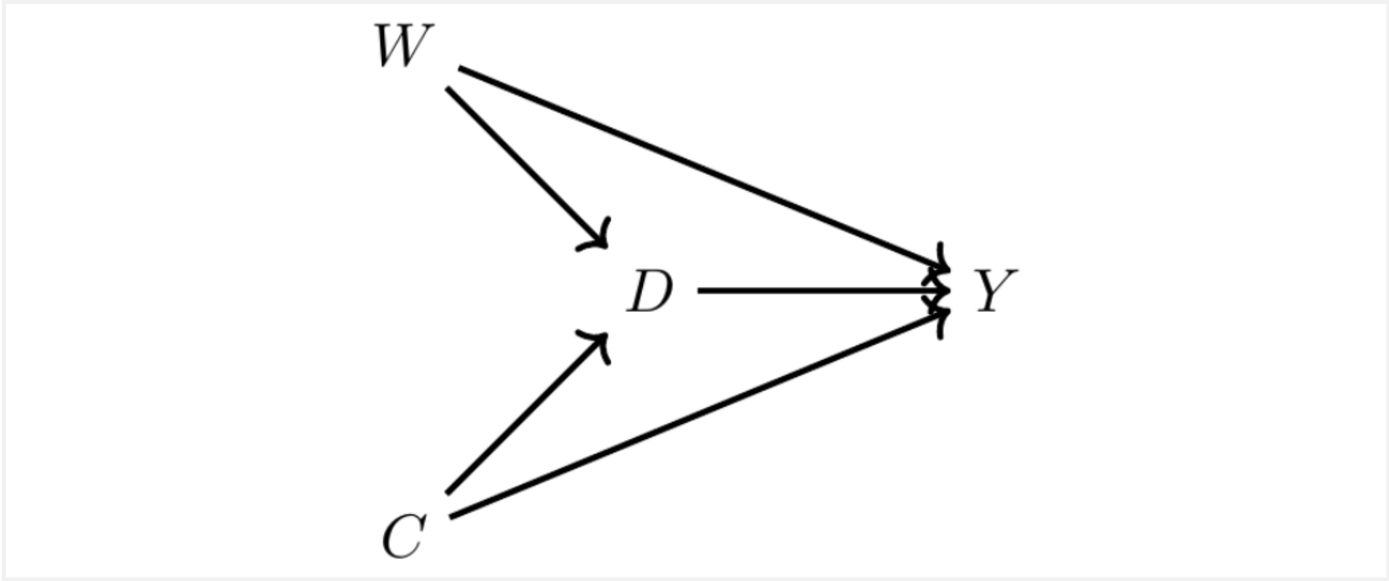
- Subclassification simply adjusts the mortality rate of cigarette smokers to have the same age distribution as the comparison groups (cigar smokers)
  - We we account for the differences in ages, cigarette smoking has the highest death rates

Table 5.4: Adjusted mortality rates using 3 age groups (Cochran 1968).

Smoking group	Canada	UK	US
Non-smokers	20.2	11.3	13.5
Cigarettes	29.5	14.8	21.2
Cigars/pipes	19.8	11.0	13.7

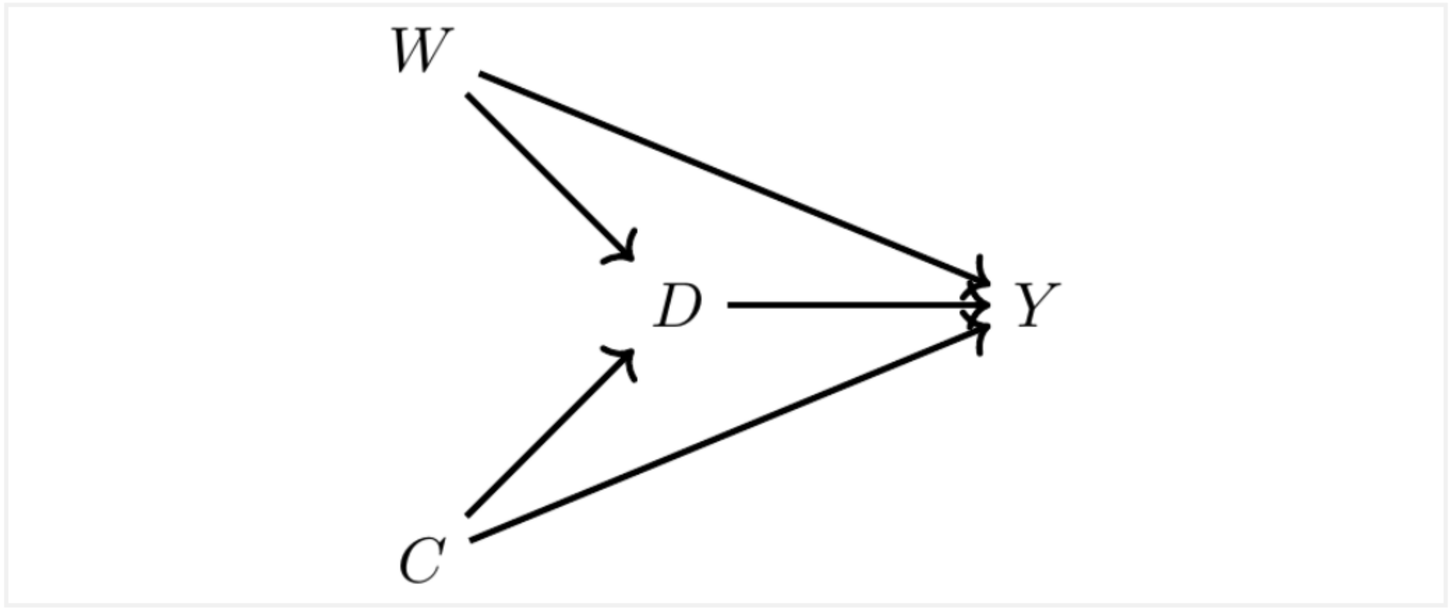
Subclassification Example

- Did first-class passengers have a better probability of surviving the sinking of the Titanic?
  - Women and children were given priority for boarding lifeboats despite wealth
- Where
  - D is first class “treatment”
  - Y is survived binary
  - W is woman
  - C is child



Subclassification Example

- Direct Pathway
  - $D \rightarrow Y$
- Indirect Pathways
  - Backdoor 1:
    - $D \leftarrow W \rightarrow Y$
  - Backdoor 2:
    - $D \leftarrow C \rightarrow Y$



Subclassification Example

- Stata Example
  - If we calculate the SDO, we can see that 1st class passengers had a very high survival probability
- 1) Stratify into 4 groups
  - Young women, older women, young men, and older men
- 2) Calculate the survival probability for each group
- 3) Calculate strata weights
  - Divide the total number of each strata group non-first-class passengers by the total number of non-first-class passengers $\frac{N_{Group-Non1st}}{N_{Total-Non1st}}$
- 4) Calculate weighted average survival rate using the strata weights

Subclassification Concluding Remarks

- Curse of Dimensionality
  - It is possible to be unable to calculate strata-specific weights as the number of strata, K, increases.
  - There may not be any observations in both treatment and control and this will violate the common support assumption
  - You can calculate the ATT if there is always control observations
- Limited use of subclassification
  - As the number of covariates grow or strata grow, we will likely run into the curse of dimensionality and spareness in observations
  - There won't be enough sample in each strata to estimate the ATE

Matching Identification Strategies

Matching Identification Strategies

- Matching Identification Strategies
  - Subclassification is not likely to be used
- There are two major types of matching identification strategies
  - Exact matching
  - Approximate matching

Exact Matching

Exact Matching

- Strengths
  - Provides a simple estimator for ATE and ATT
- Weakness
  - Matching on observable confounders, not unobservables confounders
  - Matching becomes harder as the number of covariates grow
- Assumptions
  - Conditional independence assumption (not testable)
  - Common support assumption (testable)

Exact Matching

- Abadie and Imbens (2006) provide guidance for this matching estimator
- If we fill in potential outcomes for each treatment unit by using a control unit that is closest to the treatment unit by conditioning on covariates
  - Compare a treatment unit to a control unit for the treatment unit's Y^0
- Simple Matching Estimator ATT
  - $$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D=1} (Y_i - Y_{j(i)})$$
  - Where $Y_{j(i)}$ is the closest match to Y_i for some X covariates
  - We need to fill in the missing Y^0 when $D = 1$ for the ATT

Exact Matching

- If there is more than 1 exact match for the i^{th} treatment
  - Then we take the average of the matched units
  - $$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{D=1} \left(Y_i - \left[\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(1)} \right] \right)$$
  - It is preferable for M to be small, such as $M = 2$
- Another method if M is too large is to randomly sample M for a comparison unit

Exact Matching

- We need to fill in the missing Y^0 when $D = 1$ and fill in missing Y^1 when $D = 0$ for the ATE
  - Match missing control units to treatment units
  - Match missing treatment units to control units
- $$\hat{\delta}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}(2D_i - 1)\left(Y_i - \left[\frac{1}{M}\sum_{m=1}^{M} Y_{j_m(1)}\right]\right)$$
  - Where $(2D_i - 1)$ is a switching equation:
    - When $D_i = 1$, the term is equal to 1; When $D_i = 0$ the term is equal to -1

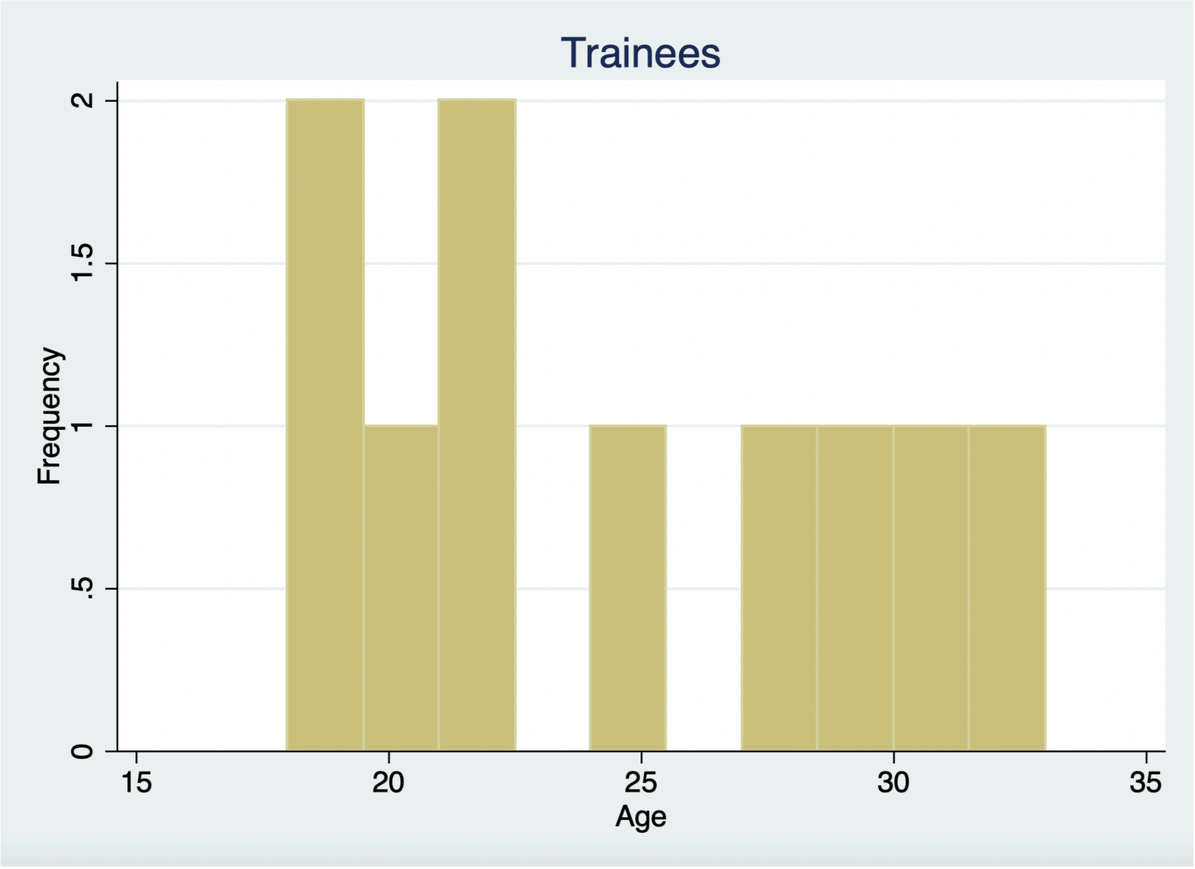
Exact Matching Example

- We have age and earnings for
  - Trainees (D=1)
  - Non-Trainees (D=0)
- SDO is -26.25
  - $11,075 vs $11,101.25
- Mean age is higher for non-trainees
  - 24.3 vs 31.95
- Calculate the ATE through 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

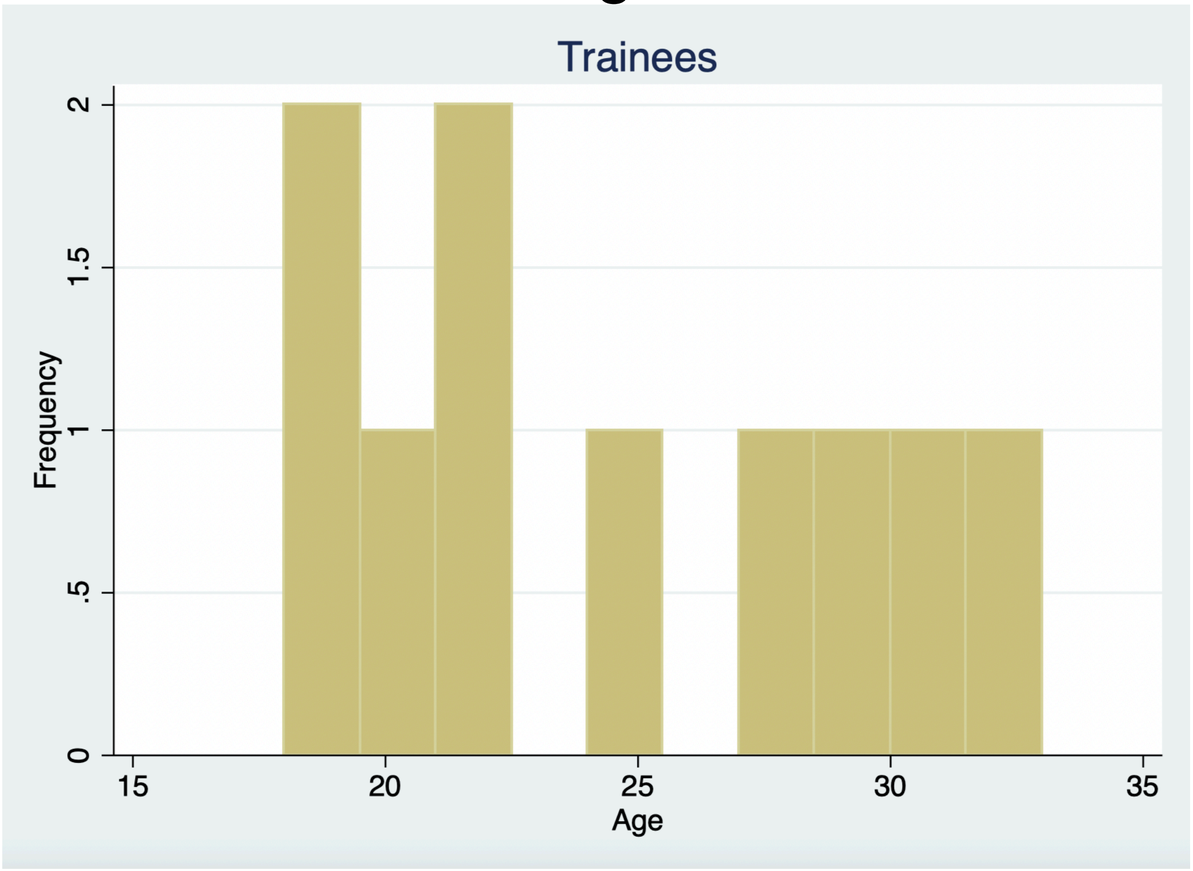
Exact Matching Example

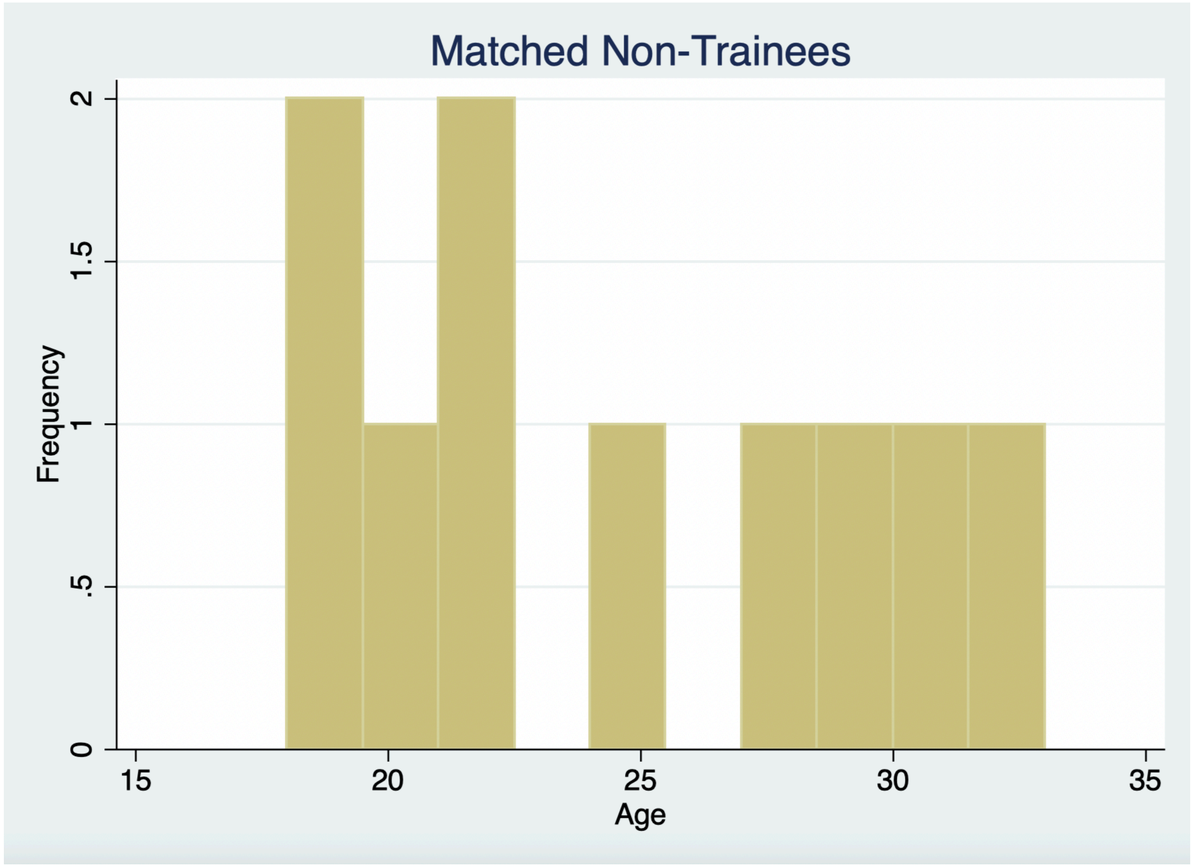
- Check the distribution of ages by trainees and non-trainees
  - Non-trainees are older and have a larger distribution than trainees



Exact Matching Example

- Matching Trainees and Non-Trainees on Age
  - Distribution of age are similar between Trainees and Matched Non-Trainees





Exact Matching Example

- Match by age
  - Compare matched
- Mean Earnings for Trainees
  - $11,075
- Mean Earnings for Matched
  - $9,380
- ATE is $1,695

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			

Exact Matching Concluding Remarks

- Exact matching estimator is simple and straightforward
- When we have a matched sample
  - We say that the groups are ***exchangeable*** when they are ***balanced***
- A dichotomous problem with exact matching
  - There is a lack of conditional independence assumption if don't match on all confounders
  - We don't have common support if we try to exact match on too many covariates


Violate
CIA Assumption
Violate
Common Support
Assumption
JAKE-CLARK.TUMBLR

Approximate Matching

Approximate Matching

- There are a few types of approximate matching strategies we will cover
  - Nearest Neighbor Matching
  - Propensity Score Matching
    - Inverse Probability Weights
    - K-Nearest Neighbor Matching
  - Coarsen Exact Matching
- Approximate matching helps when the number of covariates K grows

Nearest Neighbor Matching

- An approximate matching method such as nearest neighbor might be beneficial when the number of confounders/covariates grow
  - We want to make the i^{th} unit to the closest j^{th} unit
- We have a few methods to match for nearest neighbor
  - Euclidean Distance: $\|X_i - X_j\| = \sqrt{\sum_{i=1}^{k} (X_{ni} - X_{nj})^2}$
  - Normalized Euclidean Distance: $\|X_i - X_j\| = \sqrt{\sum_{i=1}^{k} \frac{(X_{ni} - X_{nj})^2}{\hat{\sigma}^2}}$
  - Mahalanobis Distance: $\|X_i - X_j\| = \sqrt{(X_i - X_j)' \hat{\Sigma}^{-1} 1_X (X_i - X_j)}$, where $\hat{\Sigma}_X$ is sample variance-covariance matrix

Nearest Neighbor Matching

- Nearest Neighbor Matching is limited
  - You can have more than 1 covariate, but becomes more problematic compared to other methods
  - Matching discrepancies can introduce bias
- We’ll focus on propensity score matching instead of nearest neighbor

Propensity Score Matching (PSM)

- Unlike subclassifications and nearest neighbor matching, propensity score matching (PSM) generates a scale values across multiple covariates
  - All covariates are converted into a scalar (number)
  - This eliminates the problem of dimensionality
- This method is very popular with federal evaluation researchers when randomized assignment is unavailable
  - As we will see in the TAACCCT grant evaluation, all evaluators started with randomized assignment as the initial design
  - All evaluators ended up using propensity score matching

Propensity Score Matching

- Strengths
  - Eliminates the curse of dimensionality (uses a scalar)
  - Closes backdoor pathways when conditional independence assumption is met
- Weaknesses
  - Selection on observed confounders, not unobserved
  - Can be benchmarked to RCT, but cannot substitute for RCT
- Assumptions
  - Conditional Independence Assumption (untestable)
  - Common Support Assumption (testable)

Propensity Score Matching Caveats

- You need deep institutional knowledge that the PSM is a credible identification strategy
    - You need to know that there are ***no unobserved confounders*** that undermine the identification strategy
- You may need ***trimming*** for balance
    - trimming extreme propensity scores is a way to achieve treatment-control balance
- You can benchmark a PSM estimator result to an RCT, but you cannot benchmark and RCT result to a PSM estimator

Propensity Score Matching

- Basis of Propensity Score Matching
  - PSM takes the covariates of interest and estimates a maximum likelihood model of the conditional probability of treatment
  - The logit or probit ***predicts a propensity score*** for assignment to treatment from the covariates of interest in the model
- In essence, the covariates of interest are used to calculate a single value for probability of assignment to treatment
  - Regardless if the unit actually gets treatment
  - All comparisons between treatment and control are based upon predicted propensity scores

Propensity Score Matching

- Consider two units: Unit A and Unit B
- Unit A gets the treatment
- Unit B does not get the treatment
- Their predicted propensity score is 0.6 for both Unit A and Unit B
- By random chance, Unit A gets the treatment and Unit B does not, since they both have the same conditional probability to be assigned to treatment
    - Assuming conditional independence assumption holds

Propensity Score Matching

- Implicit within Propensity Score Matching
  - Common support is an underlying assumption that needs covariate balance
  - Common support can only be tested by propensity scores
- We can check for common support so that treatment and control have overlapping propensity scores
  - If Unit A has a propensity score of 0.9 and Unit B has a propensity score of 0.1, then they do not overlap
    - The common support assumption is violated

PSM Example: Replicate NSW

- Lalonde (1986) attempts to replicate the randomized experiment of the National Supported Work (NSW) Demonstration
  - What was the impact of training on earnings?
    - Training increases earnings by $851 for AFDC women and by $886 for men
    - Note that those who joined the program earlier had a higher estimate ATE ($1,794) (Dehejia and Wahba (1999))
- To replicate the RCT without the randomized control group
  - He used control groups from the Current Population Survey (CPS) and Panel Survey of Income Dynamics (PSID)
  - His non-experimental methods performed poorly when benchmarked to the RCT
  - Why? Because $E[Y^0 \mid D = 1] \neq E[Y^0 \mid D = 0]$

PSM Example: Replicate NSW

- Covariate in T and C Lalonde (1986)
  - ***Covariate Imbalance***
  - Highly likely different Y^0
  - $E[Y^0 \mid D = 1] \neq E[Y^0 \mid D = 0]$
- Violates Independence Assumption

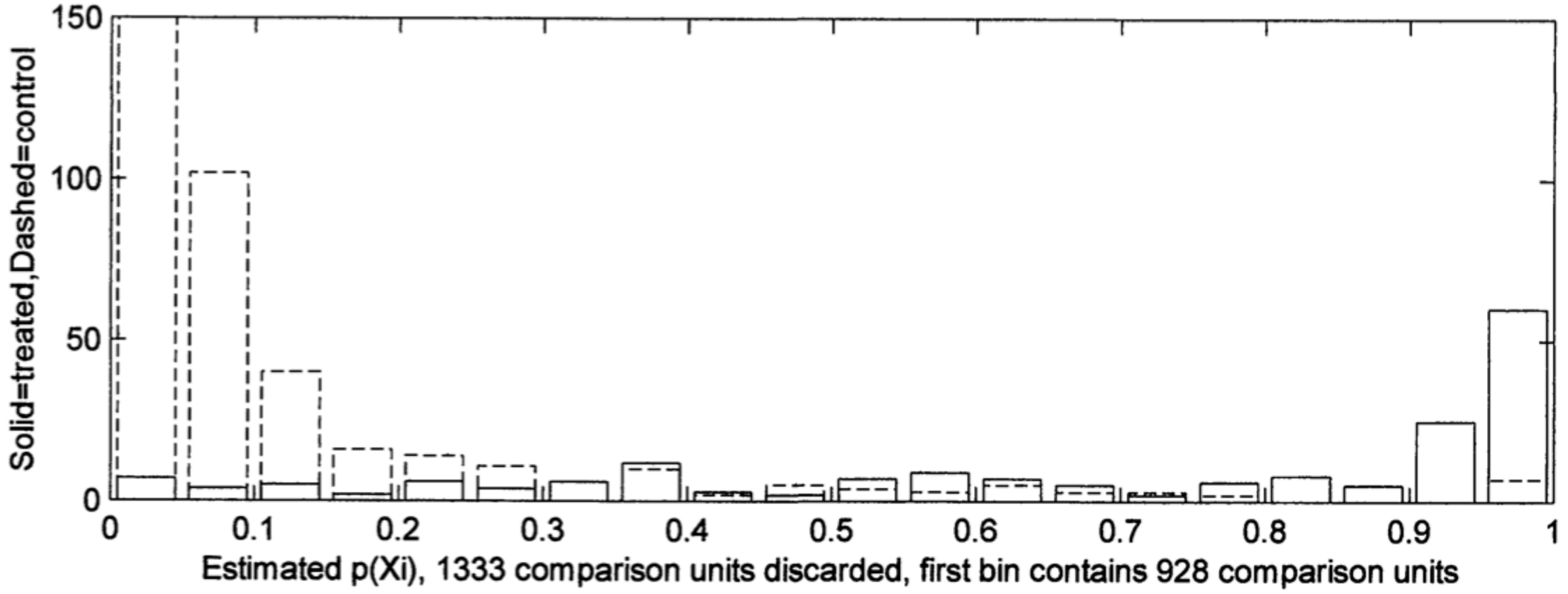
	All		CPS Controls	NSW Trainees		
	$N_c = 15,992$			$N_t = 297$		
Covariate	Mean	SD	Mean	Mean	T-statistic	Diff.
Black	0.09	0.28	0.07	0.80	47.04	-0.73
Hispanic	0.07	0.26	0.07	0.94	1.47	-0.02
Age	33.07	11.04	33.2	24.63	13.37	8.6
Married	0.70	0.46	0.71	0.17	20.54	0.54
No degree	0.30	0.46	0.30	0.73	16.27	-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

PSM Example: Replicate NSW

- Dehejia and Wahba (1999) attempts to reevaluate the NSW evaluation using propensity score matching
  - They use the same CPS and PSID samples that Lalonde (1986) uses
- Dehejia and Wahba (1999) do several steps
  - 1) Estimate the propensity score with logit
  - 2) Check for common support with histograms
  - 3) Trim for extreme values
  - 4) Recheck for common support with histograms

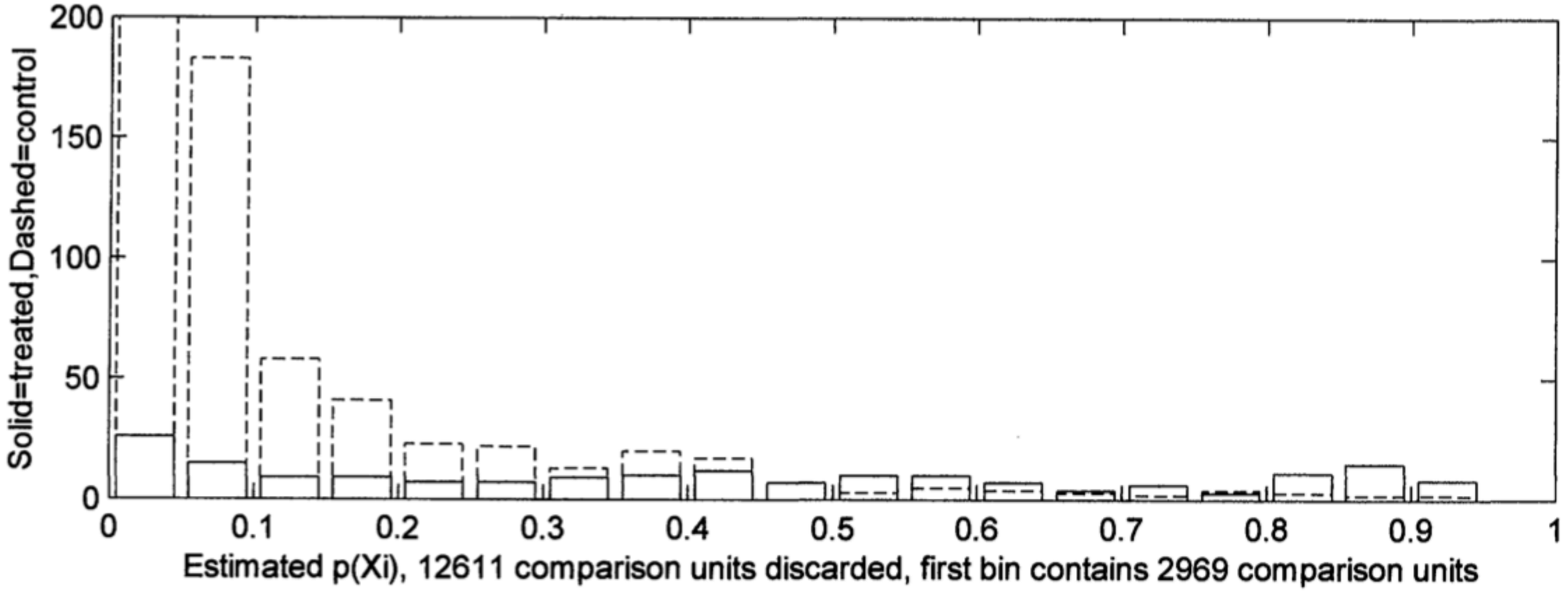
PSM Example: Replicate NSW

- Dehejia and Wahba (1999) initial common support check for PSID group



PSM Example: Replicate NSW

- Dehejia and Wahba (1999) initial common support check for CPS group



PSM Example: Replicate NSW

- Trimming is how Dehejia and Wahba (1999) get common support for NSW treatment and control (PSID and CPS)
  - 1,333 out of 2,490 PSID units and 12,611 out of 15,992 CPS units are discarded
  - Better Covariate Balance

Table 5.14: Sample Means of Characteristics for Matched Control Samples

Matched Sample	N	Age	Education	Black	Hispanic	No Degree	Married	RE74	RE75
NSW	185	25.81	10.335	0.84	0.06	0.71	0.19	2,096	1,532
PSID	56	26.39	10.62	0.86	0.02	0.55	0.15	1,794	1,126
		(2.56)	(0.63)	(0.13)	(0.06)	(0.13)	(0.13)	(0.12)	(1,406)
CPS	119	26.91	10.52	0.86	0.04	0.64	0.19	2,110	1,396
		(1.25)	(0.32)	(0.06)	(0.04)	(0.07)	(0.06)	(841)	(563)

Standard error on the difference in means with NSW sample is given in parentheses.

PSM Example: Replicate NSW

- RCT Subsample
  - $\hat{\delta}_{ATE_{RCT}} = \$1{,}672$
- PSID
  - $\hat{\delta}_{ATE_{PSM}} = \$1{,}473$
- CPS
  - $\hat{\delta}_{ATE_{PSM}} = \$1{,}616$

Table 5.13: Estimated Training Effects using Propensity Scores

	NSW T-C Earnings		Propensity Score Adjusted				
				Stratification		Matching	
Comparison group	Unadj.	Adj.	Quadratic Score	Unadj.	Adj.	Unadj.	Adj.
Experimental controls	1,794	1,672					
	(633)	(638)					
PSID-1	−15,205	731	294	1,608	1,494	1,691	1,473
	(1154)	(886)	(1389)	(1571)	(1581)	(2209)	(809)
CPS-1	−8498	972	1,117	1,713	1,774	1,582	1,616
	(712)	(550)	(747)	(1115)	(1152)	(1069)	(751)

Propensity Score Matching

- Definition of Propensity Score
  - The selection probability into treatment is conditional on the confounding variables
  - $p(X) = Pr(D = 1 \,|\, X)$
- There are two assumptions
  - Conditional Independence Assumption: $(Y^1, Y^0) \perp D \,|\, X$
  - Common Support Assumption: $0 < Pr(D = 1 \,|\, X) < 1$
- If both assumptions are satisfied: $E[\hat{\delta}(X_i)] = \delta$

Propensity Score First Step

- Estimate propensity scores with a maximum likelihood model (logit or probit)
  - $Pr(D = 1 \mid X) = F(\beta_0 + \gamma Treatment + \alpha X)$
  - Where $F = \dfrac{e}{(1 + e)}$ and X are exogenous covariates
- Estimated conditional probability to treatment
  - The propensity score estimates conditional probability to treatment **regardless of whether the unit did or did not receive treatment**

Propensity Score Second Step

- Common support is a necessary assumption that can be tested
  - It requires positive probability of being assigned to treatment
  - $0 < Pr(D = 1 \mid X) < 1$
- Common support ensures that there is sufficient overlap between treatment and control
  - There is sufficient sample in along propensity scores between treatment and control
- Methods
  - Summary statistics between treatment and control
  - Histograms between treatment and control
  - -tebalance - State command

Propensity Score Third and Fourth Steps

- Trimming will be necessary if there is imbalance in propensity scores
  - Dehejia and Wahba (1999)
  - Rule of thumb is propensity scores below 0.1 and 0.9 should be trimmed first
- Recheck the common support assumption
  - Rerun summary statistics between treatment and control
  - Rerun histograms between treatment and control
  - Rerun -tebalance-

Propensity Score Theorem

- The propensity score theorem implies ***balanced observable covariates***
- Under the Conditional Independence Assumption
  - $(Y^1, Y^0) \perp D \,|\, X$
  - Yields $(Y^1, Y^0) \perp D \,|\, p(X)$
    - Where $p(X) = Pr(D = 1 \,|\, X)$
- This means in order to achieve conditional independence, we need to condition on propensity scores
  - The only covariate you need is $p(X)$

Propensity Score Theorem

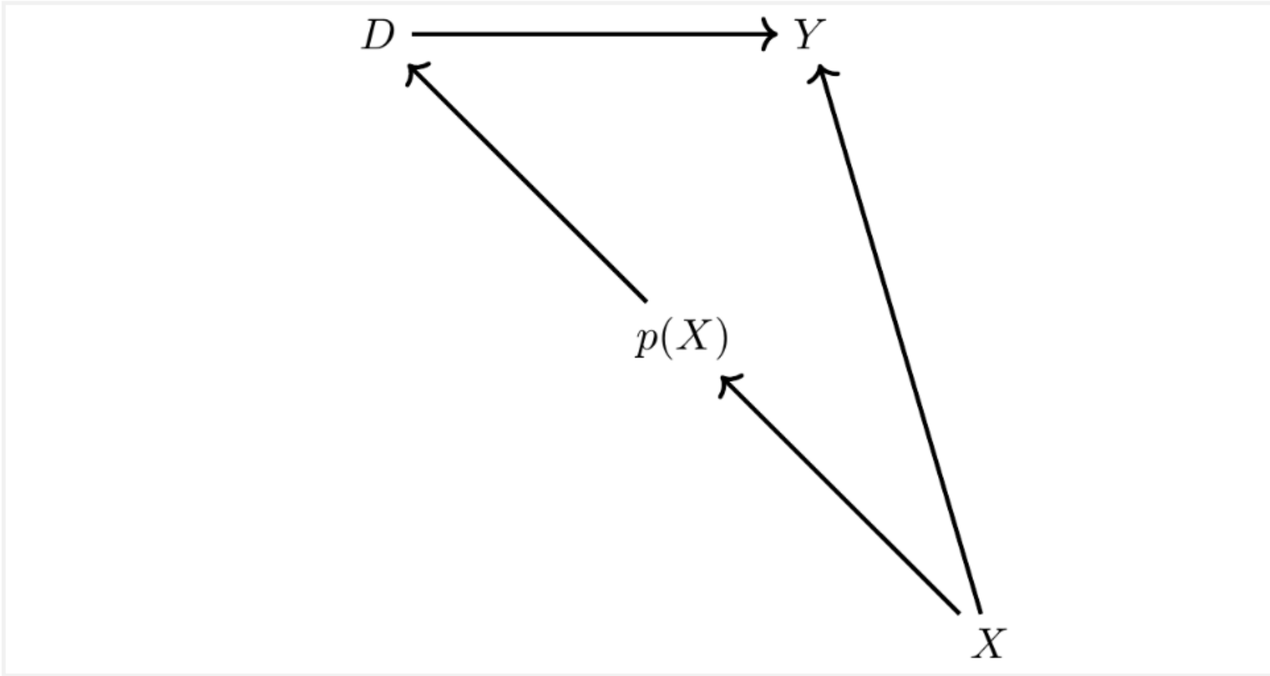
- The probability that an individual receive treatment conditional on potential outcomes and propensity score is not a function of potential outcomes
  - $Pr(D = 1 \,|\, Y^1, Y^0, p(X)) = Pr(D = 1 \,|\, p(X) = p(x)$
  - By the Conditional Independence assumption
- Propensity scores are a function of X
  - $Pr(D = 1 \,|\, X, p(X)) = Pr(D = 1 \,|\, X) = p(X)$
- Conditional on propensity scores, $D = 1$ no longer is dependent on X
  - $D \perp X \,|\, p(X)$

Propensity Score Theorem

- D and X are independent of one another conditional propensity scores
  - $D \perp X$
- We can also obtain our ***balance property*** of propensity scores from this
  - $Pr(X|D = 1,p(X)) = Pr(X|D = 0,p(X))$
- This states that covariates are the same between treatment and control conditioning on propensity scores $p(X)$

Propensity Score Theorem

- There exists two pathways between D and X
  - Indirect: $X \rightarrow Y \leftarrow D$
  - Mediated: $X \rightarrow p(x) \rightarrow D$
- Indirect is closed by collider Y
- Mediated backdoor between D and X
  - Closed by $p(X)$
- The propensity score theorem implies ***balanced observable covariates***



Propensity Score Matching Methods

- There are a few ways to calculate the ATE and ATT from propensity scores
  - Inverse Probability Weights
  - K-nearest neighbor matching
- There are other ways, but we will focus on these two methods

Inverse Probability Weights (IPW)

- Busso, DiNardo, and McCrary (2014) find that IPW is a competitive way to utilize propensity scores
- Estimate treatment effects by weighing procedure
  - A unit's propensity score is the weight of that individual's outcome (Imbens, 2000)
  - Inverse Probability Weights is not a matching method

Inverse Probability Weights (IPW)

- The average treatment effect is estimated by
  - $$\delta_{ATE} = E\left[Y^1 - Y^0] = E[Y \cdot \frac{D - p(X)}{p(X)(1 - p(X)}\right]$$
- The average treatment on the treated is estimated by
  - $$\delta_{ATT} = E[Y^1 - Y^0 \,|\, D = 1] = \frac{1}{Pr(D = 1)} \cdot E\left[Y \cdot \frac{D - p(X)}{1 - p(X)}\right]$$

Inverse Probability Weights (IPW)

- Getting estimates of ATE and ATT from samples is a two-step procedure
- Estimated ATE
  - $$\hat{\delta}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \frac{D - \hat{p}(X_i)}{\hat{p}(X_i) \cdot (1 - \hat{p}(X_i))}$$
- Estimated ATT
  - $$\hat{\delta}_{ATT} = \frac{1}{N_T} \sum_{i=1}^{N} Y_i \cdot \frac{D - \hat{p}(X_i)}{(1 - \hat{p}(X_i))}$$

Calculating Standard Errors for IPW

- Bootstrapping is a methodology for estimating the variance by Efron (1979
  - Use to estimating standard errors around our $\hat{\delta}_{ATE}$ and $\hat{\delta}_{ATT}$
- To bootstrap
  - We take repeated random “draws” (with replacement) from our original data
  - Estimate $\hat{\delta}_{ATE}$ and $\hat{\delta}_{ATT}$
  - Repeat for 1,000 or 10,000 times, we get a distribution of $\hat{\delta}_{ATE}$ and $\hat{\delta}_{ATT}$
- This can be easily implemented in Stata as an option

Normalized Inverse Probability Weights (IPW)

- As an alternative to IPW, you can use normalized IPW
  - This helps when extreme values of propensity scores
  - Instead of $\frac{1}{N}$ for weights
  - Use weights that are normalized by the sum of propensity scores for treated and control groups
  - Can easily be implemented with -teffects- or -ipw- in Stata
- $$\hat{\delta}_{ATT} = \left[\sum_{i=1}^{N} \frac{Y_i D_i}{\widehat{p}}\right] / \left[\sum_{i=1}^{N} \frac{D_i}{\widehat{p}}\right] - \left[\sum_{i=1}^{N} \frac{Y_i(1-D_i)}{(1-\widehat{p})}\right] / \left[\sum_{i=1}^{N} \frac{(1-D_i)}{(1-\widehat{p})}\right]$$

K-Nearest Neighbor with Propensity Scores

- A popular alternative to IPW is K-nearest neighbor matching
  - This is a matching strategy that uses propensity scores for matching
  - Find the nearest k-neighbors within a particular radius of the treatment unit's propensity score
  - Then average the k-neighbors for a comparison unit to the treatment unit
- This matching process is easily implemented in Stata
  - -teffects psmatch- with the nn(k) option

K-Nearest Neighbor with Propensity Scores

- King and Nelson (2019)
  - Note that trimming can amplify bias with K-NN matching strategy
  - The more balanced the data are, or become through trimming, the more propensity score will degrade in inferences
- Regardless, this K-NN method and IPW are very popular matching strategies
- Estimator after K-NN matches
  - $\hat{\delta}_{ATT} = \frac{1}{N_T}(Y_i - Y_{j(i)})$ where $Y_{j(i)}$ is our K-NN

Coarsen Exact Matching

Coarsen Exact Matching

- Iacus, King, and Porro (2012) introduced a new matching strategy called coarsen exact matching
  - Exact matches on continuous variable, such as income, age, become impossible
- The premise
  - Exact matching on continuous variables becomes possible if we generate bins from the continuous variables (or coarsen)
  - Example: we coarsen age into bins, such as 0-10, 11-17, 18-22, 24-29, etc.

Coarsen Exact Matching Steps

- Create a copy of a set of variables X
- Coarsen the set of covariates X with user defined bins or CEM algorithm
- Assign each unit a place in the strata
- Assign these strata to the uncoarsen data
- Drop any observation whose stratum does not contain at least 1 treatment and control observation
- Add weights for the stratum size
- Analyze the strata without matching

Coarsen Exact Matching

- Trade off
  - Larger bins means more coarsen data and more covariate imbalance
  - You need to be upfront about your trimming
- Benefit of CEM
  - Coarsen exact matching uses monotonic imbalance bounding
  - Ex ante (before hand) choices by the user determine the imbalance
  - The user controls how much imbalance there is

Coarsen Exact Matching

- It is called imbalance bounding, since imbalance is bounded between 0 and 1
- Measuring imbalance
- $$L1(f, g) = \frac{1}{2} \sum_{l_1,\ldots,l_k} \left| f_{l_1,\ldots,l_k} - g_{l_1,\ldots,l_k} \right|$$
  - Where f and g are relative frequencies of treatment and control units
  - Perfect balance $L1 = 0$ and perfect imbalance $L1 = 1$

CEM Balance

- Estimate with CEM
- Age and Age-Squared
  - The highest imbalance
- The $\hat{\delta}_{ATE} = \$2{,}152$
  - It is higher than RCT

Covariate	L1	Mean	Min.	25%	50%	75%	Max.
age	.08918	.55337	1	1	0	1	0
agesq	.1155	21.351	33	35	0	49	0
agecube	.05263	626.9	817	919	0	1801	0
school	6.0e-16	$-2.3e-14$	0	0	0	0	0
schoolsq	5.4e-16	$-2.8e-13$	0	0	0	0	0
married	1.1e-16	$-1.1e-16$	0	0	0	0	0
nodegree	4.7e-16	$-3.3e-16$	0	0	0	0	0
black	4.7e-16	$-8.9e-16$	0	0	0	0	0
hispanic	7.1e-17	$-3.1e-17$	0	0	0	0	0
re74	.06096	42.399	0	0	0	0	-94.801
re75	.03756	-73.999	0	0	0	-222.85	-545.65
u74	1.9e-16	$-2.2e-16$	0	0	0	0	0
u75	2.5e-16	$-1.1e-16$	0	0	0	0	0
interaction1	.06535	425.68	0	0	0	0	-853.21

Matching Concluding Remarks

- Subclassification and Exact Matching will likely be difficult matching strategies to implement
  - Curse of Dimensionality
- If our matching strategy satisfies the Condition Independence Assumption and Common Support
  - Our PSM strategy will identify the causal effect
- This is difficult since we cannot prove CIA is satisfied
  - We can benchmark our result to an RCT but we cannot know how good our estimated treatment effects are without an RCT