Matching and the Propensity Score

2020 AEA Continuing Education Program
Mastering Mostly Harmless Econometrics

Alberto Abadie
MIT

Adjustment techniques for observational studies

- Alternatives to regression:
  - Subclassification
  - Matching
  - Propensity Score Methods
- What should we match on? A brief introduction to DAGs.

Covariates and outcomes

Definition (Predetermined Covariates)

Variable X is predetermined with respect to the treatment D (also called "pretreatment") if for each individual i, $X_{0i} = X_{1i}$, ie. the value of X_i does not depend on the value of D_i. Such characteristics are called *covariates*.

- Does not imply that X and D are independent
- Predetermined variables are often time invariant (sex, race, etc.), but time invariance is not necessary

Definition (Outcomes)

Those variables, Y, that are (possibly) not predetermined are called outcomes (for some individual i, $Y_{0i} \neq Y_{1i}$)

In general, one should not condition on outcomes, because this may induce bias

Identification in randomized experiments

Randomization implies:

$$(Y_1, Y_0) \text{ independent of } D, \quad \text{or} \quad (Y_1, Y_0) \perp\!\!\!\perp D.$$

Therefore:

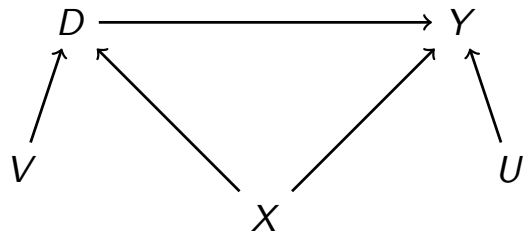
$$\begin{aligned} E[Y|D=1] - E[Y|D=0] &= E[Y_1|D=1] - E[Y_0|D=0] \\ &= E[Y_1] - E[Y_0] \\ &= E[Y_1 - Y_0]. \end{aligned}$$

Also, we have that

$$E[Y_1 - Y_0] = E[Y_1 - Y_0|D=1].$$

The nature of confounding

- Confounding may arise from common causes in observational studies:



- X is a confounder, V and U are not.
- Conditional on X there is no confounding.
- Correlation between Y and D conditional on X is reflective of the effect of D on Y. That is:

$$(Y_1, Y_0) \perp\!\!\!\perp D|X.$$

Identification under selection on observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identification Result

Given selection on observables we have

$$\begin{aligned} E[Y_1 - Y_0|X] &= E[Y_1 - Y_0|X, D = 1] \\ &= E[Y|X, D = 1] - E[Y|X, D = 0] \end{aligned}$$

Therefore, under the common support condition:

$$\begin{aligned} \alpha_{ATE} &= E[Y_1 - Y_0] = \int E[Y_1 - Y_0|X]\, dP(X) \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0]\right) dP(X) \end{aligned}$$

Identification under selection on observables

Identification Assumption

1. $(Y_1, Y_0) \perp\!\!\!\perp D|X$ *(selection on observables)*
2. $0 < \Pr(D = 1|X) < 1$ *with probability one (common support)*

Identification Result

Similarly,

$$\begin{aligned} \alpha_{ATET} &= E[Y_1 - Y_0|D = 1] \\ &= \int \left(E[Y|X, D = 1] - E[Y|X, D = 0]\right) dP(X|D = 1) \end{aligned}$$

To identify α_{ATET} the selection on observables and common support conditions can be relaxed to:

- $Y_0 \perp\!\!\!\perp D|X$
- $\Pr(D = 1|X) < 1$ *(with* $\Pr(D = 1) > 0$*)*

The subclassification estimator

The identification result is:

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

Assume X takes on K different cells $\{X^1, ..., X^k, ..., X^K\}$.Then, the analogy principle suggests the following estimators:

$$\widehat{\alpha}_{ATE} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N^k}{N}\right); \ \widehat{\alpha}_{ATET} = \sum_{k=1}^{K} \left(\bar{Y}_1^k - \bar{Y}_0^k\right) \cdot \left(\frac{N_1^k}{N_1}\right)$$

- N^k is # of obs. and N_1^k is # of treated obs. in cell k
- $\bar{Y}_1^k$ is mean outcome for the treated in cell k
- $\bar{Y}_0^k$ is mean outcome for the untreated in cell k

Subclassification and the "curse of dimensionality"

- Subclassification becomes unfeasible with many covariates
- Assume we have k covariates and divide each of them into 3 coarse categories (e.g., age could be "young", "middle age" or "old", and income could be "low", "medium" or "high").
- The number of subclassification cells is 3^k. For $k = 10$, we obtain $3^{10} = 59049$
- Many cells may contain only treated or untreated observations, so we cannot use subclassification
- Subclassification is also problematic if the cells are "too coarse". But using "finer" cells worsens the curse of dimensionality problem: e.g., using 10 variables and 5 categories for each variable we obtain $5^{10} = 9765625$

Matching

- We could also estimate α_{ATET} by constructing a comparison sample of untreated units with the same characteristics as the sample of treated units.
- This can be easily accomplished **matching** treated and untreated units with the same characteristics.

Matching: An ideal 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

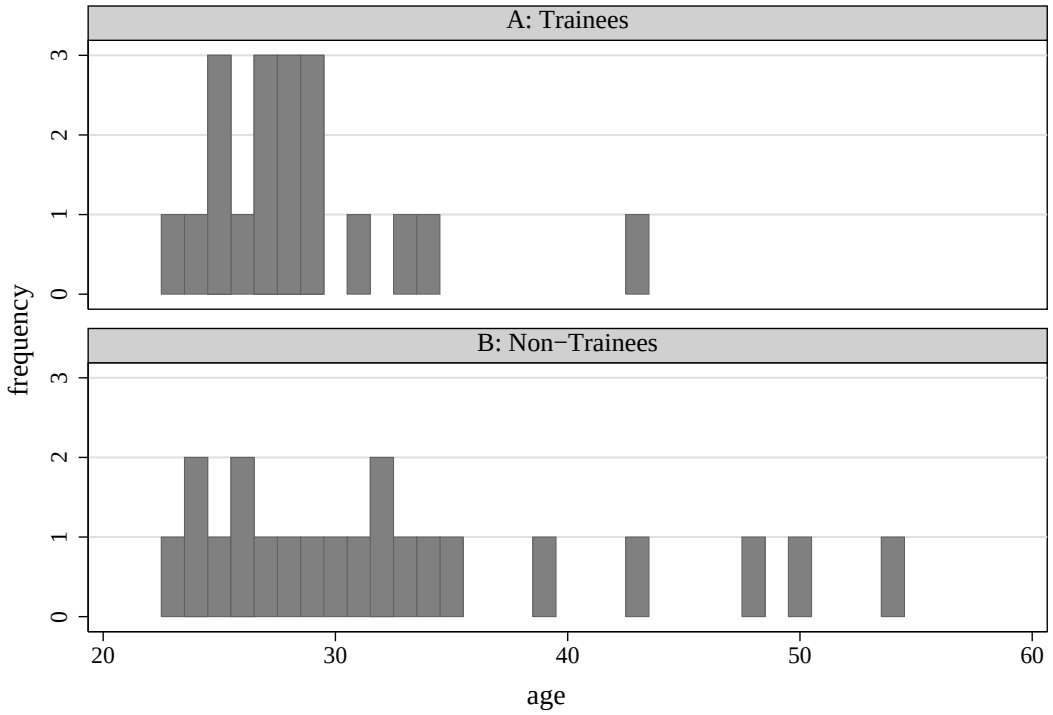
Matching: An ideal example

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

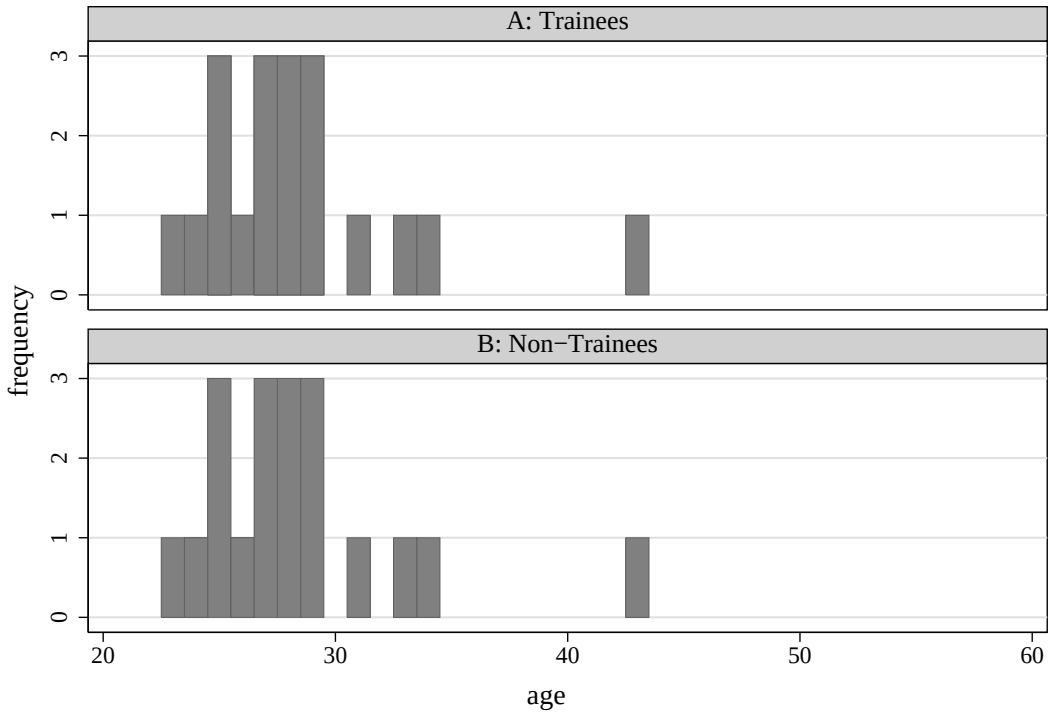
Non-Trainees		
unit	age	earnings
1	43	20900
2	50	31000
3	30	21000
4	27	9300
5	54	41100
6	48	29800
7	39	42000
8	28	8800
9	24	25500
10	33	15500
11	26	400
12	31	26600
13	26	16500
14	34	24200
15	25	23300
16	24	9700
17	29	6200
18	35	30200
19	32	17800
20	23	9500
21	32	25900
Average:	33	20724

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



Treatment effect estimates

Difference in average earnings between trainees and non-trainees:

- Before matching

$$16426 - 20724 = -4298$$

- After matching:

$$16426 - 13982 = 2444$$

Matching

Perfect matches are often not available. In that case, a matching estimator of α_{ATET} can be constructed as:

$$\widehat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{D_i=1} \left(Y_i - Y_{j(i)} \right)$$

where $Y_{j(i)}$ is the outcome of an untreated observation such that $X_{j(i)}$ is the **closest** value to X_i among the untreated observations.

We can also use the average for M closest matches:

$$\widehat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{D_i=1} \left\{ Y_i - \left(\frac{1}{M} \sum_{m=1}^{M} Y_{j_m(i)} \right) \right\}$$

Works well when we can find good matches for each treated unit, so M is usually small (typically, $M = 1$ or $M = 2$).

Matching

We can also use matching to estimate α_{ATE}. In that case, we match in both directions:

1. If observation i is treated, we impute Y_{0i} using untreated matchs, $\{Y_{j_1(i)}, \ldots, Y_{j_M(i)}\}$
2. If observation i is untreated, we impute Y_{1i} using treated matchs, $\{Y_{j_1(i)}, \ldots, Y_{j_M(i)}\}$

The estimator is:

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

Matching and the curse of dimensionality

- When we match multiple variables we need to define a norm, $\|\cdot\|$, to measure **matching discrepancies**, $\|X_i - X_{j(i)}\|$ (see appendix for usual norms)
- Matching discrepancies $\|X_i - X_{j(i)}\|$ tend to increase with k, the dimension of X
- Matching discrepancies converge to zero. But they converge very slowly if k is large
- Mathematically, it can be shown that $\|X_i - X_{j(i)}\|$ converges to zero at the same rate as $\frac{1}{N^{1/k}}$
- It is difficult to find good matches in large dimensions: you need many observations if k is large

Matching: Bias

$$\widehat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{D_i=1} (Y_i - Y_{j(i)}),$$

where $X_i \simeq X_{j(i)}$ and $D_{j(i)} = 0$. Let

$$\begin{aligned} \mu_0(x) &= E[Y|X = x, D = 0] = E[Y_0|X = x], \\ \mu_1(x) &= E[Y|X = x, D = 1] = E[Y_1|X = x], \\ Y_i &= \mu_{D_i}(X_i) + \varepsilon_i. \end{aligned}$$

Then,

$$\begin{aligned} \widehat{\alpha}_{ATET} - \alpha_{ATET} &= \frac{1}{N_1} \sum_{D_i=1} (\mu_1(X_i) - \mu_0(X_i) - \alpha_{ATET}) \\ &+ \frac{1}{N_1} \sum_{D_i=1} (\varepsilon_i - \varepsilon_{j(i)}) \\ &+ \frac{1}{N_1} \sum_{D_i=1} (\mu_0(X_i) - \mu_0(X_{j(i)})). \end{aligned}$$

Matching: Bias

We hope that we can apply a Central Limit Theorem and

$$\sqrt{N_1}(\widehat{\alpha}_{ATET} - \alpha_{ATET})$$

converges to a Normal distribution with zero mean. However,

$$E[\sqrt{N_1}(\widehat{\alpha}_{ATET} - \alpha_{ATET})] = E[\sqrt{N_1}(\mu_0(X_i) - \mu_0(X_{j(i)}))|D = 1].$$

Now, if k is large:

- $\Rightarrow$ The difference between X_i and $X_{j(i)}$ converges to zero very slowly
- $\Rightarrow$ The difference $\mu_0(X_i) - \mu_0(X_{j(i)})$ converges to zero very slowly
- $\Rightarrow$ $E[\sqrt{N_1}(\mu_0(X_i) - \mu_0(X_{j(i)}))|D = 1]$ may not converge to zero!
- $\Rightarrow$ $E[\sqrt{N_1}(\widehat{\alpha}_{ATET} - \alpha_{ATET})]$ may not converge to zero!

$\Rightarrow$ Bias is often an issue when we match in many dimensions

Matching: Reducing bias

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

1. The matching discrepancies are observed. We can always check in the data how well we are matching the covariates.
2. For $\widehat{\alpha}_{ATET}$ we can always make the matching discrepancies small by using a large reservoir of untreated units to select the matches (that is, by making N_0 large).
3. If the matching discrepancies are large, so we are worried about potential biases, we can apply bias correction techniques.
4. Partial solution: Propensity score methods (to come).

Matching with bias correction

Each treated observation contributes

$$\mu_0(X_i) - \mu_0(X_{j(i)})$$

to the bias.

Bias-corrected matching:

$$\widehat{\alpha}^{BC}_{ATET} = \frac{1}{N_1} \sum_{D_i=1} \Big((Y_i - Y_{j(i)}) - (\widehat{\mu}_0(X_i) - \widehat{\mu}_0(X_{j(i)})) \Big)$$

where $\widehat{\mu}_0(x)$ is an estimate of $E[Y|X = x, D = 0]$ (e.g., OLS).

Matching bias: Implications for practice

Bias arises because of the effect of large matching discrepancies on $\mu_0(X_i) - \mu_0(X_{j(i)})$. To minimize matching discrepancies:

1. Use a small M (e.g., $M = 1$). Large 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.

Propensity score

The **propensity score** is defined as the selection probability conditional on the confounding variables: $p(X) = P(D = 1|X)$.

The selection on observables identification assumption is:

1. $(Y_1, Y_0) \perp\!\!\!\perp D \mid X$ (selection on observables)
2. $0 < \Pr(D = 1|X) < 1$ (common support)

Rosenbaum and Rubin (1983) proved that selection on observables implies:

$$(Y_1, Y_0) \perp\!\!\!\perp D \mid p(X)$$

$\Rightarrow$ conditioning on the propensity score is enough to have independence between the treatment indicator and the potential outcomes

$\Rightarrow$ substantial dimension reduction in the matching variables!

Matching on the propensity score

Because of the Rosenbaum-Rubin result, if $(Y_1, Y_0) \perp\!\!\!\perp D|X$, then

$$E[Y_1 - Y_0|p(X)] = E[Y|D = 1, p(X)] - E[Y|D = 0, p(X)]$$

This motivates a two step procedure to estimate causal effects under selection on observables:

1. estimate the propensity score $p(X) = P(D = 1|X)$ (e.g., using logit or probit regression)
2. do matching or subclassification on the estimated propensity score

Proof of the Rosenbaum and Rubin (1983) result

Assume that $(Y_1, Y_0) \perp\!\!\!\perp D \,|\, X$. Then:

$$\begin{aligned} P(D = 1|Y_1, Y_0, p(X)) &= E[D|Y_1, Y_0, p(X)] \\ &= E\left[E[D|Y_1, Y_0, X]|Y_1, Y_0, p(X)\right] \\ &= E\left[E[D|X]|Y_1, Y_0, p(X)\right] \\ &= E\left[p(X)|Y_1, Y_0, p(X)\right] \\ &= p(X) \end{aligned}$$

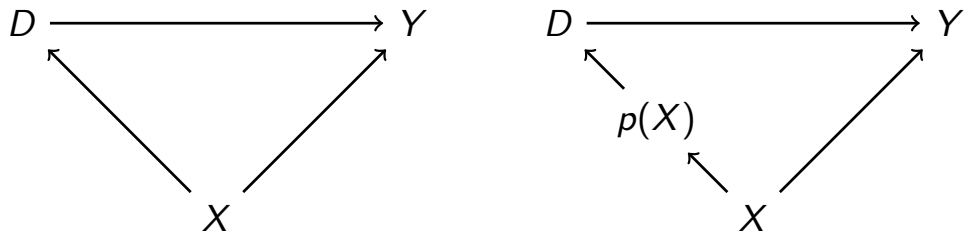
Using a similar argument, we obtain

$$\begin{aligned} P(D = 1|p(X)) &= E[D|p(X)] = E[E[D|X]|p(X)] \\ &= E[p(X)|p(X)] = p(X) \end{aligned}$$

$\Rightarrow P(D = 1|Y_1, Y_0, p(X)) = P(D = 1|p(X))$

$\Rightarrow (Y_1, Y_0) \perp\!\!\!\perp D \,|\, p(X)$

Propensity score: Balancing property



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

$$D \perp\!\!\!\perp X \mid p(X).$$

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

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

$\Rightarrow$ conditional on the propensity score, the distribution of the covariates is the same for treated and non-treated.

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

$$P(X|D = 1, \widehat{p}(X)) = P(X|D = 0, \widehat{p}(X))$$

Matching estimators: Large sample distribution

- Matching estimators have a Normal distribution in large samples (provided that the bias is small).
- Abadie and Imbens (2006, 2012) provide standard errors formulas for estimators that match on X.
- Abadie and Imbens (2016) provide standard errors formulas for estimators that match on $\widehat{p}(X)$.
- The bootstrap does not work in general.

Weighting on the propensity score (IPW)

Weighting estimators that use the propensity score ("Inverse Probability Weighting") are based on the following result: If $Y_1, Y_0 \perp\!\!\!\perp D|X$, then

$$\alpha_{ATE} = E\left[Y \frac{D - p(X)}{p(X)(1 - p(X))}\right]$$
$$\alpha_{ATET} = \frac{1}{P(D=1)} E\left[Y \frac{D - p(X)}{1 - p(X)}\right]$$

To prove this results notice that:

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

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

Weighting on the propensity score

$$\alpha_{ATE} = E\left[Y \frac{D - p(X)}{p(X)(1 - p(X))}\right]$$
$$\alpha_{ATET} = \frac{1}{P(D=1)} E\left[Y \frac{D - p(X)}{1 - p(X)}\right]$$

The analogy principle suggests a two step estimator:

1. Estimate the propensity score: $\widehat{p}(X)$
2. Use estimated score to produce analog estimators:

$$\widehat{\alpha}_{ATE} = \frac{1}{N} \sum_{i=1}^{N} Y_i \frac{D_i - \widehat{p}(X_i)}{\widehat{p}(X_i)(1 - \widehat{p}(X_i))}$$
$$\widehat{\alpha}_{ATET} = \frac{1}{N_1} \sum_{i=1}^{N} Y_i \frac{D_i - \widehat{p}(X_i)}{1 - \widehat{p}(X_i)}$$

Weighting on the propensity score

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N} Y_i \frac{D_i - \widehat{p}(X_i)}{\widehat{p}(X_i)(1-\widehat{p}(X_i))}$$

$$\widehat{\alpha}_{ATET} = \frac{1}{N_1}\sum_{i=1}^{N} Y_i \frac{D_i - \widehat{p}(X_i)}{1-\widehat{p}(X_i)}$$

Several improvements and variants have been proposed (e.g., normalizing the weights so that they sum to one, Imbens 2004).

Standard errors:

- We need to adjust the s.e.'s for first-step estimation of $p(X)$
- Parametric $p(X)$: Newey & McFadden (1994)
- Non-parametric $p(X)$: Newey (1994), Hirano, Imbens, and Ridder (2003)
- Or bootstrap the entire two-step procedure

Doubly robust estimators

Combine propensity-score based and regression based estimation. Estimators that only need the propensity score or the regression function to be correctly specified (Bang and Robins, 2005).

$$\widehat{\alpha}_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i\left(Y_i - \widehat{\mu}_1(X_i)\right)}{\widehat{p}(X_i)} + \widehat{\mu}_1(X_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\left(\frac{(1-D_i)\left(Y_i - \widehat{\mu}_0(X_i)\right)}{1-\widehat{p}(X_i)} + \widehat{\mu}_0(X_i)\right).$$

Doubly robust estimators, and more generally locally robust estimators (Chernozhukov et al., 2016) have appealing properties in terms of bias and in terms of inference after model selection in the first step.

Abadie and Imbens (2011)

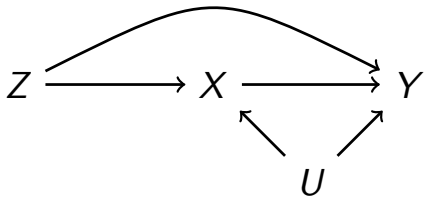
Table 2. Experimental and nonexperimental estimates for the NSW data

	$M=1$		$M=4$		$M=16$		$M=64$		$M=2490$	
	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)	Est.	(SE)
Panel A:										
Experimental estimates										
Covariate matching	1.22	(0.84)	1.99	(0.74)	1.75	(0.74)	2.20	(0.70)	1.79	(0.67)
Bias-adjusted cov matching	1.16	(0.84)	1.84	(0.74)	1.54	(0.75)	1.74	(0.71)	1.72	(0.68)
Pscore matching	1.43	(0.81)	1.95	(0.69)	1.85	(0.69)	1.85	(0.68)	1.79	(0.67)
Bias-adjusted pscore matching	1.22	(0.81)	1.89	(0.71)	1.78	(0.70)	1.67	(0.69)	1.72	(0.68)
Regression estimates										
Mean difference	1.79	(0.67)								
Linear	1.72	(0.68)								
Quadratic	2.27	(0.80)								
Weighting on pscore	1.79	(0.67)								
Weighting and linear regression	1.69	(0.66)								
Panel B:										
Nonexperimental estimates										
Simple matching	2.07	(1.13)	1.62	(0.91)	0.47	(0.85)	−0.11	(0.75)	−15.20	(0.61)
Bias-adjusted matching	2.42	(1.13)	2.51	(0.90)	2.48	(0.83)	2.26	(0.71)	0.84	(0.63)
Pscore matching	2.32	(1.21)	2.06	(1.01)	0.79	(1.25)	−0.18	(0.92)	−1.55	(0.80)
Bias-adjusted pscore matching	3.10	(1.21)	2.61	(1.03)	2.37	(1.28)	2.32	(0.94)	2.00	(0.84)
Regression estimates										
Mean difference	−15.20	(0.66)								
Linear	0.84	(0.88)								
Quadratic	3.26	(1.04)								
Weighting on pscore	1.77	(0.67)								
Weighting and linear regression	1.65	(0.66)								

NOTE: The outcome is earnings in 1978 in thousands of dollars.

What to match on: A brief introduction to DAGs

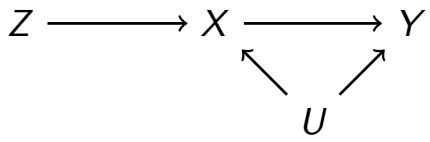
A **Directed Acyclic Graph** (**DAG**) is a set of nodes (vertices) and directed edges (arrows) with no directed cycles.



- Nodes represent variables.
- Arrows represent direct causal effects ("direct" means not mediated by other variables in the graph).
- A **causal DAG** must include:
  1. All direct causal effects among the variables in the graph
  2. All common causes (even if unmeasured) of any pair of variables in the graph

Some DAG concepts

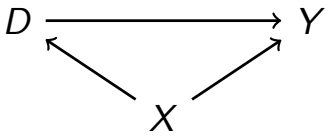
In the DAG:



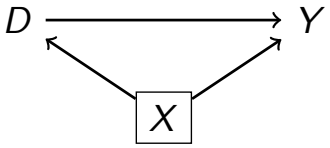
- U is a **parent** of X and Y.
- X and Y are **descendants** of Z.
- There is a **directed path** from Z to Y.
- There are two **paths** from Z to U (but no directed path).
- X is a **collider** of the path $Z \rightarrow X \leftarrow U$.
- X is a **noncollider** of the path $Z \rightarrow X \rightarrow Y$.

Confounding

- Confounding arises when the treatment and the outcome have common causes.



  The association between D and Y does not only reflect the causal effect of D on Y.
- Confounding creates **backdoor paths**, that is, paths starting with incoming arrows. In the DAG we can see a backdoor path from D to Y ($D \leftarrow X \rightarrow Y$).
- However, once we "block" the backdoor path by conditioning on the common cause, X, the association between D and Y is only reflective of the effect of D on Y.



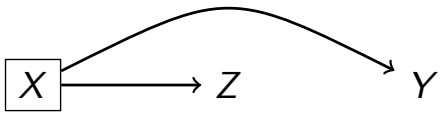
Blocked paths

A path is blocked if and only if:

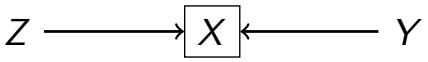
- It contains a noncollider that has been conditioned on,
- Or, it contains a collider that has not been conditioned on and has no descendants that have been conditioned on.

Examples:

1. Conditioning on a noncollider blocks a path:



2. Conditioning on a collider opens a path:



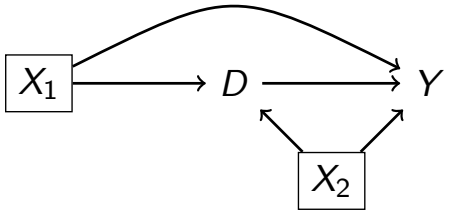
3. Not conditioning on a collider (or its descendants) leaves a path blocked:

Backdoor criterion

- Suppose that:
  - D is a treatment,
  - Y is an outcome,
  - $X_1, \ldots, X_k$ is a set of covariates.
- Is it enough to match on $X_1, \ldots, X_k$ in order to estimate the causal effect of D on Y? Pearl's **Backdoor Criterion** provides sufficient conditions.
- Backdoor criterion: $X_1, \ldots, X_k$ satisfies the backdoor criterion with respect to (D, Y) if:
  1. No element of $X_1, \ldots, X_k$ is a descendant of D.
  2. All backdoor paths from D to Y are blocked by $X_1, \ldots, X_k$.
- If $X_1, \ldots, X_k$ satisfies the backdoor criterion with respect to (D, Y), then matching on $X_1, \ldots, X_k$ identifies the causal effect of D on Y.

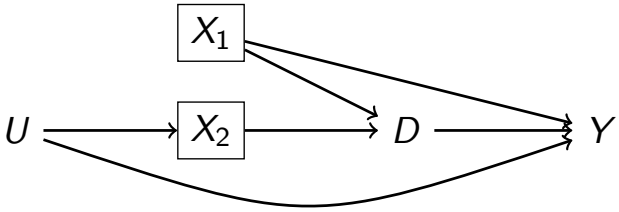
Implications for practice

- **Matching on all common causes is sufficient:** There are two backdoor paths from D to Y.



Conditioning on X_1 and X_2 blocks the backdoor paths.

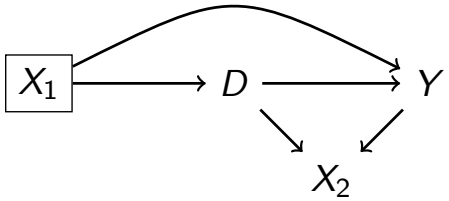
- **Matching may work even if not all common causes are observed:** U and X_1 are common causes.



Conditioning on X_1 and X_2 is enough.

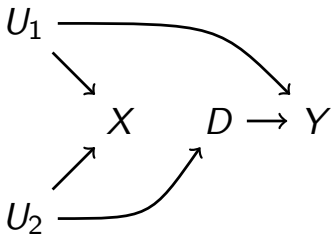
Implications for practice (cont.)

- **Matching on an outcome may create bias:** There is only one backdoor path from D to Y.



Conditioning on X_1 blocks the backdoor path. Conditioning on X_2 would open a path!

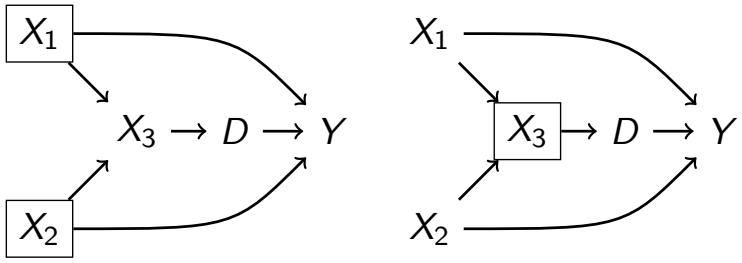
- **Matching on all pretreatment covariates is not always the answer:** There is one backdoor path and it is closed.



No confounding. Conditioning on X would open a path!

Implications for practice (cont.)

- **There may be more than one set of conditioning variables that satisfy the backdoor criterion:**



  - Conditioning on the common causes, X_1 and X_2, is sufficient, as always.
  - But conditioning on X_3 only also blocks the backdoor paths.

Appendix: Matching Distance Metric

Matching: Distance metric

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 need to define a **distance metric** to measure "closeness". The usual **Euclidean distance** is:

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

$\Rightarrow$ The Euclidean distance is not invariant to changes in the scale of the X's.

$\Rightarrow$ For this reason, we often use alternative distances that are invariant to changes in scale.

Matching: Distance metric

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

$$\widehat{V} = \begin{pmatrix} \widehat{\sigma}_1^2 & 0 & \cdots & 0 \\ 0 & \widehat{\sigma}_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \widehat{\sigma}_k^2 \end{pmatrix}.$$

Notice that, the normalized Euclidean distance is equal to:

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

$\Rightarrow$ Changes in the scale of X_{ni} affect also $\widehat{\sigma}_n$, and the normalized Euclidean distance does not change.

Matching: Distance metric

Another popular scale-invariant distance is the **Mahalanobis distance**:

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

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

We can also define arbitrary distances:

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

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