MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 3: Selection on Observables 1

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2024

Experiments and Observational Studies

Randomized experiments are called the gold standard for (internal validity of) causal inference.

But we cannot (should not?) always randomize!

Enter observational studies: Designs where the assignment mechanism is not known or not under researcher's control.

Goal is to design studies such that we believe causal effects are still identified, and understand and evaluate the assumptions underpinning these designs.

Begin with selection on observables – an assumption-heavy design that provides the ground work for much more.

Lecture Roadmap

1. Covariates
2. Identification: Potential Outcomes
3. Identification: Graphical
4. Estimation: Subclassification

Pre-Treatment Covariates

Definition (pre-treatment covariate)

Any variable X that is predetermined with respect to the treatment D such that the value of X_i for each unit i does not depend on the value of D_i.

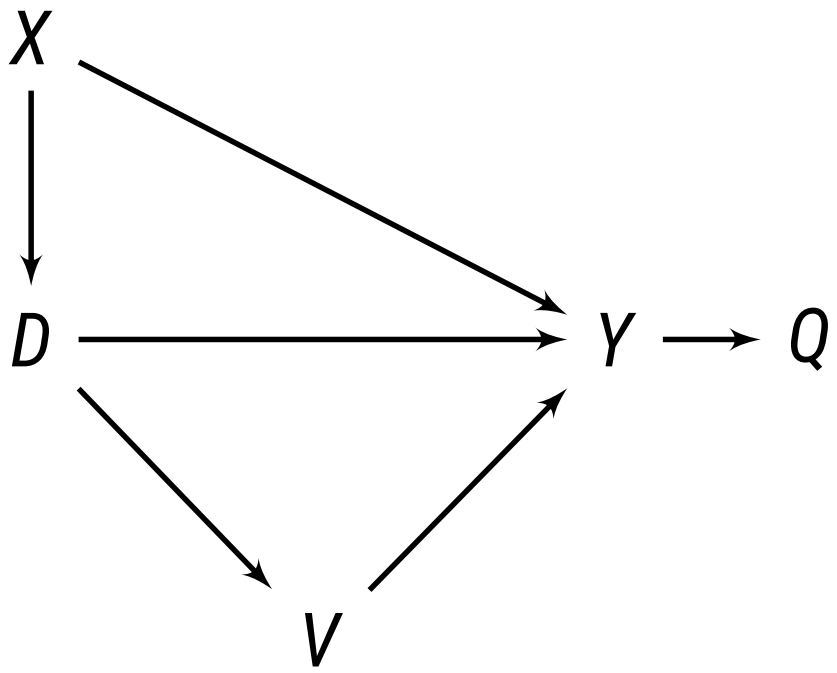
This implies that there are no potential outcomes X_{0i} and X_{1i} with respect to this treatment D, just one value X_i, taken as fixed for the purposes of our analysis.

X and D may still be associated if the treatment assignment for D is associated with or causally affected by X.

X may include characteristics that are immutable (e.g. age) or they may be causally affected by other things (e.g. income).

X may include baseline (pre-treatment) measures of Y.

Pre-Treatment vs. Post-Treatment Covariates



From this perspective, post-treatment covariates are descendents of D. They may be direct descendents (e.g. V above) or indirect descendents e.g Q above.

Identification Assumptions

In randomized experiments, D_i satisfies independence (or ignorability):

$$(Y_{i0}, Y_{i1}) \perp\!\!\!\perp D_i$$

What if we cannot assume independence? Instead, we might assume:

1. The conditional ignorability (CI) (a.k.a exogeneity, independence) assumption:

$$(Y_{i0}, Y_{i1}) \perp\!\!\!\perp D_i \mid X_i = x \quad \text{for any} \quad x \in \mathcal{X}$$

Read: Among units with identical values of X_i, D_i is "as-if" random.

2. The common support (a.k.a positivity, overlap) assumption:

$$0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \text{for any} \quad x \in \mathcal{X}$$

Read: With any value of X_i, i could have received treatment or control.

Identification Result for ATE

Previously we considered identification with population difference in means:

$$\hat{\tau} \;=\; \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$$

Consider instead the difference in population regression functions:

$$\hat{\tau}_{CATE}(x) \;=\; \mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]$$

Result: Under our two assumptions, ATE is nonparametrically identified as:

$$\begin{aligned} \tau_{ATE} &= \mathbb{E}[\hat{\tau}_{CATE}(X_i)] \\ &= \int (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]) f(x) dx \end{aligned}$$

where the first $\mathbb{E}$ is taken with respect to the distribution of X_i, $f(x)$.

Identification Result for ATT

ATT is also nonparametrically identified under the conditional ignorability and common support assumptions as:

$$\tau_{ATT} \quad = \quad \mathbb{E}[\hat{\tau}_{CATE}(X_i) \mid D_i = 1]$$

where $\mathbb{E}$ is taken with respect to the distribution of X_i given $D_i = 1$.

However, the identification assumptions may be relaxed for the ATT:

1. $(Y_{i0}) \perp\!\!\!\perp D_i \mid X_i = x$
2. $\Pr(D_i = 1 \mid X_i = x) < 1$ (a.k.a "weak overlap")

Does $\tau_{ATE} = \tau_{ATT}$ necessarily hold when conditional ignorability holds? No!

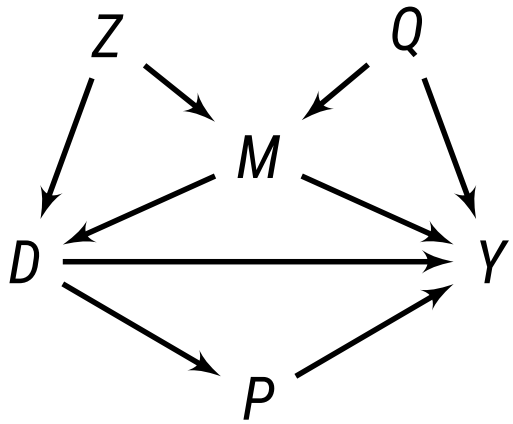
Why? $\mathbb{E}[\hat{\tau}(x) \mid D_i = 1] \neq \mathbb{E}[\hat{\tau}(x)]$ when D_i is not unconditionally random.

Blocked Paths

Definition (blocked paths)

A set of nodes $\{\mathbf{S}\}$ blocks a path $\boldsymbol{p}$ if either

1. $\boldsymbol{p}$ contains at least one *arrow-emitting node* in $\mathbf{S}$, or
2. $\boldsymbol{p}$ contains at least one *collision node* that is outside $\mathbf{S}$ and has no descendant in $\mathbf{S}$.

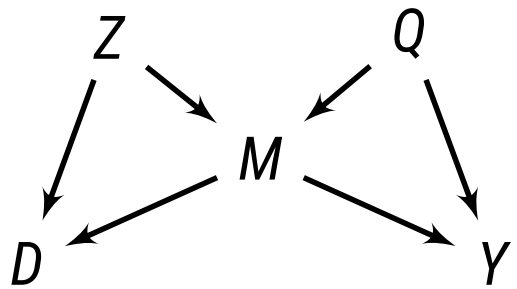


The path $D \rightarrow P \rightarrow Y$ is blocked by $\{P\}$
The path $D \leftarrow M \rightarrow Y$ is blocked by $\{M\}$
The path $D \leftarrow Z \rightarrow M \rightarrow Y$ is blocked by $\{M\}$ or $\{Z\}$ or $\{Z, M\}$
The path $D \leftarrow Z \rightarrow M \leftarrow Q \rightarrow Y$ is blocked by $\{Z\}$ or $\{Q\}$ or $\{\varnothing\}$

d-Separation

Definition (*d*-separation)

If **S** blocks all paths from D to Y, then **S** *d*-separates D and Y.
If **S** *d*-separates D and Y, then $D \perp\!\!\!\perp Y \mid \mathbf{S}$.



D and Y are *d*-separated by $\{Z, M\}$ or $\{Q, M\}$ or $\{Z, Q, M\}$.

The Back-Door Criterion for Causal Identification

The graphical concept of ***d*-separation** corresponds to the statistical concept of conditional independence.

This leads to a powerful theorem for causal inference (Pearl, 2000):

Theorem (back-door criterion)

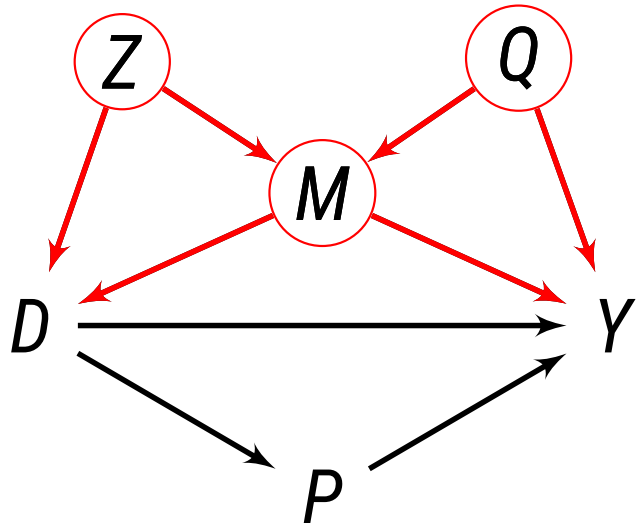
A set **S** *is sufficient for adjustment to identify the causal effect of* ***X*** *on* ***Y*** *if:*

1. *No element of* **S** *is a descendant of* ***X****, and*
2. *The elements of* **S** *block all back-door paths from* ***X*** *to* ***Y***

Note: Pearl (2000) also gives us a front-door criterion for identification, but it is hard to find effective examples in the real world, so we won't dive deeper now. See Glynn & Kashin (2017) and Bellemare et al. (2024).

Identification via Back-Door Criterion: Example

Consider again our DAG:



What conditioning set(s) identify the total effect of D on Y?

$\{Z, M\}$ or $\{M, Q\}$ or $\{Z, Q, M\}$. Why?

Only $\{M\}$ opens a back-door path due to the collider M.

Only $\{Z, Q\}$ (or either alone) leaves a back-door path open.

Good Control, Bad Control?

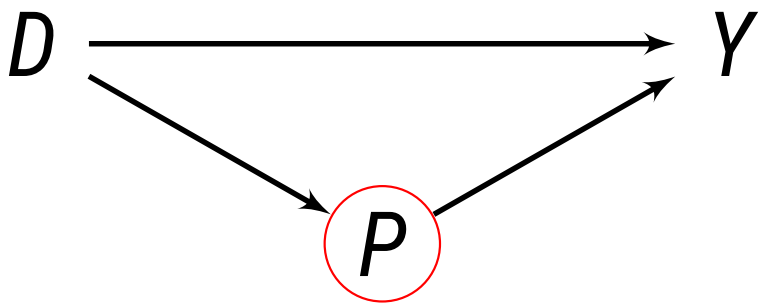
The graphical framework provides some insights that aren't always apparent when using potential outcomes. One set of insights relates to whether particular controls are "good", "bad", or "neutral" in terms of identification and efficiency.

Cinelli et al. (2022) provide a survey of multiple example models that demonstrate cases of good, bad, and neutral controls. Very useful reference!

Good controls tend to be those that block backdoor paths (establishing identification). Good controls can also be those that improve precision (regardless of identification).

Note: These insights assume our DAG is (close to) correct!

Good, Bad, or Neutral?



This is a bad control, a case of overcontrol (or post-treatment) bias. Why?

The total effect (τ_{ATE}) is given by the combination of $D \rightarrow Y$ and $D \rightarrow P \rightarrow Y$. By adjusting for P we instead get the controlled direct effect: $D \rightarrow Y$.

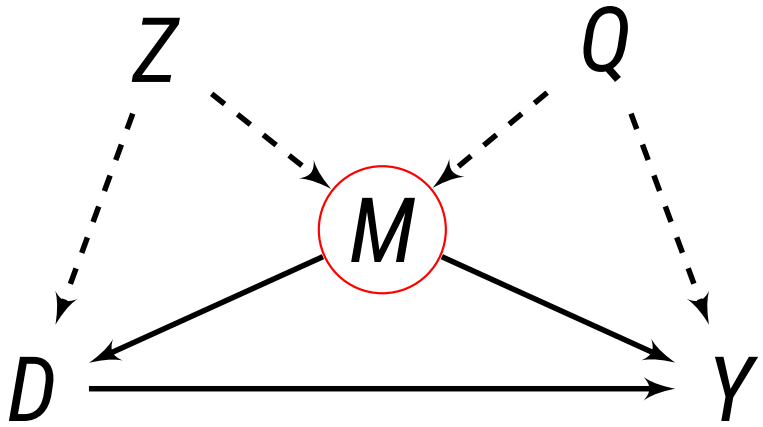
This can be a useful quantity, but it requires our DAG to be correct! See e.g. Acharya et al. (2016) for more.

Good, Bad, or Neutral?

This is a neutral control that may improve efficiency. Why?

In this DAG, Q affects Y, but is unrelated to D. By conditioning on Q we control away noise in Y. All that remains is variation in Y that is induced by D, so efficiency may improve.

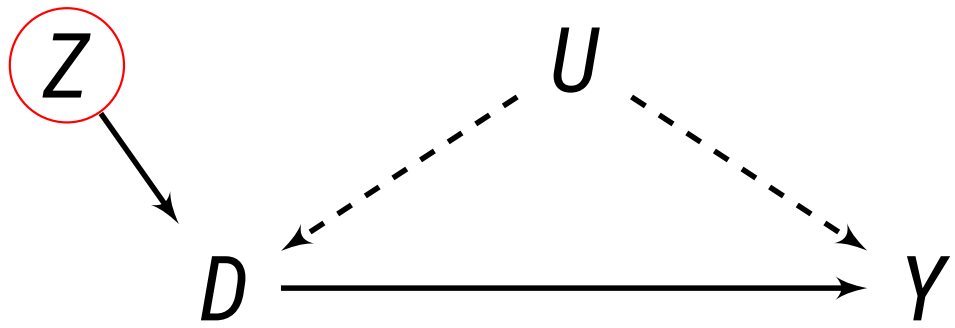
Good, Bad, or Neutral?



This is a bad control, a case of M-bias. Why?

As we saw earlier, adjusting for M we open a back-door path that was otherwise blocked! In this DAG, no observable conditioning set identifies $D \rightarrow Y$.

Good, Bad, or Neutral?



This is a bad control, a case of bias amplification. Why?

In this DAG, Z sets D exogenously (to Y). By conditioning on Z we control away exogenous variation in D. All that remains is endogenous variation in D, and so the confounding effect of U is amplified.

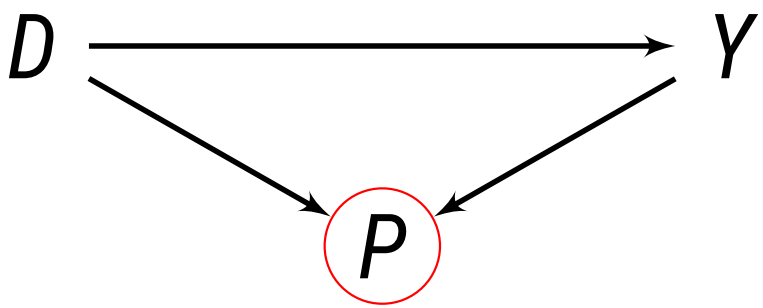
Good, Bad, or Neutral?

This is a neutral control that may harm efficiency. Why?

In this DAG, Z sets D exogenously (to Y). By conditioning on Z we do not threaten identification, but we control away "inferentially helpful" variation in D (and by implication Y).

General rule of thumb #1: Controlling for predictors of D is much less helpful (often harmful) than controlling for predictors of Y.

Good, Bad, or Neutral?



This is a bad control, a case of collider stratification bias. Why?

In this DAG, *D* and *Y* both set *P*. By conditioning on *P* we open a back-door path.

General rule of thumb #2: Don't condition on descendents of *D* (post-treatment covariates). There are some instances where this can be appropriate, but they are few and far between.

From Identification to Estimation

We now consider four broad approaches for estimating causal estimands under conditioning:

1. Subclassification
2. Matching
3. Weighting
4. Regression

These are sometimes different, sometimes identical, depending on the situation

Consider subclassification today, the rest next week.

Subclassification with Discrete Covariates

Recall our SOO identification result. If X_i is all discrete, the identification results can be rewritten as:

$$\tau_{ATE} = \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]) \Pr(X_i = x)$$

$$\tau_{ATT} = \sum_{x \in \mathcal{X}} (\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]) \Pr(X_i = x \mid D_i = 1)$$

That is, the ATE is given by:

1. Grouping units into strata (or cells) defined by the values of X_i.
2. For each stratum, calculating the difference in means of Y_i.
3. Taking weighted average of (2), where weights are the prop. of units per strata.

Similarly, the ATT is given by:

1 - 2. Same as for ATE.

3. Calculating the weighted average of (2), with weights equal to the proportions of units in the strata within the treatment group.

Subclassification Estimators

This result can be easily translated into two subclassification estimators for a given sample:

$$\hat{\tau}_{ATE} = \sum_{j=1}^{M} (\overline{Y}_{1j} - \overline{Y}_{0j}) \frac{n_j}{n}$$

$$\hat{\tau}_{ATT} = \sum_{j=1}^{M} (\overline{Y}_{1j} - \overline{Y}_{0j}) \frac{n_{1j}}{n_1}$$

where

$$\begin{aligned} M &= \text{\# of strata} \\ n_j &= \text{\# of units in cell } j \\ n_{1j} &= \text{\# of treated units in cell } j \\ \overline{Y}_{dj} &= \text{mean outcome for units with } D_i = d \text{ in cell } j \end{aligned}$$

Canonical Example: Smoking and Mortality (Cochran 1968)

TABLE 1

DEATH RATES PER 1,000 PERSON-YEARS

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

Example: Smoking and Mortality (Cochran 1968)

TABLE 2

MEAN AGES, YEARS

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

Subclassification Example

To control for differences in age, we would like to compare different smoking-habit groups with the same age distribution.

Subclassification allows us to do just this:

1. for each country, divide each group into different age subgroups
2. calculate death rates within age subgroups
3. average within age subgroup death rates using fixed weights (e.g., number of cigarette smokers)

Subclassification Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for pipe smokers?

$$15 \cdot (11/40) + 35 \cdot (13/40) + 50 \cdot (16/40) = 35.5$$

Subclassification Example

	Death Rates Pipe Smokers	# Pipe-Smokers	# Non-Smokers
Age 20 - 50	15	11	29
Age 50 - 70	35	13	9
Age + 70	50	16	2
Total		40	40

What is the average death rate for pipe smokers if they had the same age distribution as non-smokers?

$$15 \cdot (29/40) + 35 \cdot (9/40) + 50 \cdot (2/40) = 21.2$$

Subclassification Example

TABLE 3

ADJUSTED DEATH RATES USING 3 AGE GROUPS

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

Subclassification by Age ($J = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimate of the ATE of smoking on death rate?

$$\hat{\tau}_{ATE} = (28 - 24) \cdot \frac{10}{20} + (22 - 16) \cdot \frac{10}{20} = 5$$

Subclassification by Age ($J = 2$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old	28	24	3	10
Young	22	16	7	10
Total			10	20

What is the subclassification estimate of the ATT of smoking on death rate?

$$\hat{\tau}_{ATT} = (28 - 24) \cdot \frac{3}{10} + (22 - 16) \cdot \frac{7}{10} = 5.4$$

Subclassification by Age and Gender ($J = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate of the ATE of smoking on death rate?

Not identified! Why?

Lack of common support means one of our missing potential outcomes is not estimable (without additional assumptions)

Subclassification by Age and Gender ($J = 4$)

X_j	Death Rate Smokers	Death Rate Non-Smokers	# Smokers	# Obs.
Old, Male	28	22	3	7
Old, Female		24	0	3
Young, Male	21	16	3	4
Young, Female	23	17	4	6
Total			10	20

What is the subclassification estimate of the ATT of smoking on death rate?

$$\begin{aligned}\hat{\tau}_{ATT} &= (28-22)\cdot\frac{3}{10}+(21-16)\cdot\frac{3}{10}+(23-17)\cdot\frac{4}{10}\\ &= 5.1\end{aligned}$$