

# MY457/MY557: Causal Inference for Observational and Experimental Studies

## Week 3: Selection on Observables 1

Daniel de Kadt

Department of Methodology  
LSE

Winter Term 2026

# 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 then evaluate the **assumptions** underpinning these designs.

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

1 Covariates

2 Identification: Potential Outcomes

3 Identification: Graphical

# Pre-Treatment Covariates

## Pre-treatment covariate:

Any variable  $\mathbf{X}$  that is predetermined with respect to the treatment  $\mathbf{D}$  such that the value of  $\mathbf{X}_i$  for each unit  $i$  does not depend on the value of  $\mathbf{D}_i$ .

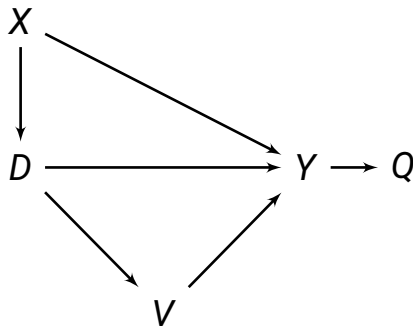
Note: This implies that there are **no potential outcomes  $\mathbf{X}_{0i}$  and  $\mathbf{X}_{1i}$**  with respect to this treatment  $\mathbf{D}$ , just one value  $\mathbf{X}_i$ , taken as fixed for the purposes of our analysis.

$\mathbf{X}$  and  $\mathbf{D}$  may still be associated if the treatment assignment for  $\mathbf{D}$  is associated with or causally affected by  $\mathbf{X}$ .

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

$\mathbf{X}$  may include baseline (pre-treatment) measures of  $\mathbf{Y}$ .

## Pre-Treatment vs. Post-Treatment Covariates



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

1 Covariates

2 Identification: Potential Outcomes

3 Identification: Graphical

# Identification Assumptions

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

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i$$

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

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

$$(Y_{0i}, Y_{1i}) \perp\!\!\!\perp D_i \mid X_i = x \quad \text{for any } x \in \mathcal{X}$$

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:

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

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

# Identification Result for ATE

1. Let's first reason about CATEs:

$$\begin{aligned}\tau_{CATE}(x) &= \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x] \\ &= \mathbb{E}[Y_{1i} \mid X_i = x] - \mathbb{E}[Y_{0i} \mid X_i = x] \\ &= \mathbb{E}[Y_{1i} \mid D_i = 1, X_i = x] - \mathbb{E}[Y_{0i} \mid D_i = 0, X_i = x] \quad \because \text{CI} \\ &= \underbrace{\mathbb{E}[Y_i \mid D_i = 1, X_i = x] - \mathbb{E}[Y_i \mid D_i = 0, X_i = x]}_{\text{conditional difference-in-means}}\end{aligned}$$

2. Now, we reason about our true estimand, the ATE:

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= \int \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = x] d\mathbf{P}(X_i) \quad \because \text{common support} \\ &= \int (\mathbb{E}[Y_i \mid D_i = 1, X_i] - \mathbb{E}[Y_i \mid D_i = 0, X_i]) d\mathbf{P}(X_i) \\ &\quad \because \text{proof 1 above.}\end{aligned}$$



# Identification Result for ATE

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

$$\begin{aligned}\tau_{ATE} &= \mathbb{E}[\hat{\tau}_{CATE}(\mathbf{X}_i)] \\ &= \int (\mathbb{E}[Y_i \mid D_i = 1, \mathbf{X}_i] - \mathbb{E}[Y_i \mid D_i = 0, \mathbf{X}_i]) d\mathbf{P}(\mathbf{X}_i)\end{aligned}$$

where the first  $\mathbb{E}$  is taken with respect to the distribution of  $\mathbf{X}_i$  for all  $i$ ,  $\mathbf{P}(\mathbf{X}_i)$ .

Read: The ATE is identified as a weighted average of **differences** in the **population regression function** for every value of  $\mathbf{X}_i$  for all  $i$ ...

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

... where the **weights** correspond to the probability of  $\mathbf{X}_i = \mathbf{x}$ .

## Identification Result for ATT

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

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

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

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

1.  $(Y_{0i}) \perp\!\!\!\perp D_i \mid X_i = x$
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**.

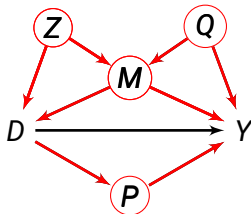
- 1 Covariates
- 2 Identification: Potential Outcomes
- 3 Identification: Graphical

# Blocked Paths

## Blocked paths:

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

1.  $p$  contains at least one *arrow-emitting node* in  $\mathbf{S}$ , or
2.  $p$  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  $\{M\}$

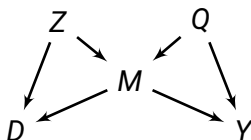
The path  $D \leftarrow Z \rightarrow M \rightarrow Y$  is blocked by  $\{M\}$  or  $\{Z\}$  or  $\{Z, M\}$

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

# $d$ -Separation

## $d$ -separation:

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

**Back-door criterion** (Pearl, 2000):

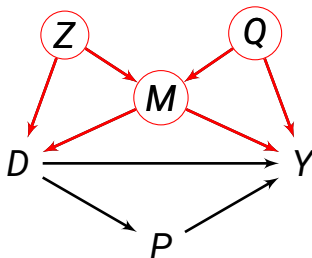
A set **S** is sufficient for adjustment to identify the causal effect of **X** on **Y** if:

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 not  $\{M\}$ ? Only  $\{M\}$  opens a back-door path due to the collider  $M$ .

Similarly, 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**.

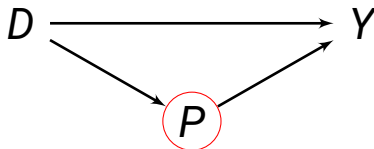
Cinelli et al. (2022) provide a survey of multiple example models that demonstrate cases of good, bad, and neutral controls. Very useful reference!

**Good controls** tend to be those that block backdoor paths (facilitating **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** ( $\tau_{ATE}$ ) is a combination of  $D \rightarrow Y$  and  $D \rightarrow P \rightarrow Y$ . By adjusting for  $P$  we instead identify the **controlled direct effect**:  $D \rightarrow Y$ .

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

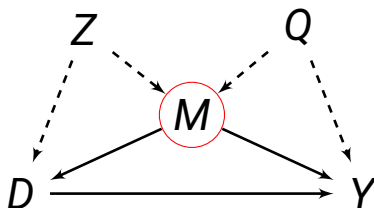
## Good, Bad, or Neutral?



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.

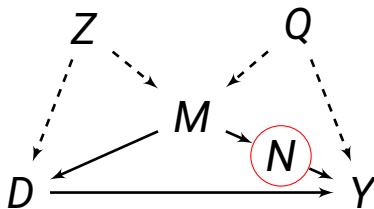
## Good, Bad, or Neutral?



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

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

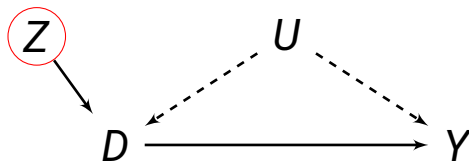
## Good, Bad, or Neutral?



This is a bad control and still a case of **M-bias**. Why?

By adjusting for a **descendant** of  $M$  we “partially” adjust for  $M$ , **opening** the back-door path that was **otherwise blocked**. As before, no observed conditioning set identifies  $D \rightarrow Y$  here.

## Good, Bad, or Neutral?



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

In this DAG,  $Z$  sets  $D$  **exogenously** (w.r.t.  $Y$ ). 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?

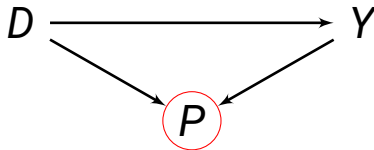


This is a neutral control that may **harm efficiency**. Why?

In this DAG,  $Z$  sets  $D$  **exogenously** (w.r.t.  $Y$ ). 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 than controlling for predictors of  $Y$  (and can even be harmful).

## 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 descendants of  $D$  (post-treatment covariates). There are **some instances** where this can be appropriate, but they are few and far between.

# What Next? From Identification to Estimation

Today we learned the **identification assumptions** required for **selection on observables** as a causal research design.

Also learned about **different types of covariates** and some of the (often non-obvious) implications of conditioning on them.

Next week we will study four broad approaches to **estimating causal estimands under selection on observables**:

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