Lecture 5: Causal Inference - Part 1

LSE ME314: Introduction to Data Science and Machine Learning (https://github.com/me314-lse)

2025-07-21

Daniel de Kadt

Slides last updated: 21 July 2025

Where Did We Begin?

Where Did We Begin?

Imagine you're a member of Netflix's c-suite. What kind of questions might you have about Squid Game?

- ➜ Did viewers like Squid Game?
- ➜ Which viewers most liked Squid Game?
- ➜ What types of viewers engaged with Squid Game?
- ➜ Did Squid Game increase Netflix viewership?
- ➜ What was the $ value of Squid Game?
- ➜ Who should we advertise or push Squid Game to?
- ➜ How many $ should we spend on advertising Squid Game?
- ➜ Should we invest in Season 4?
- ➜ What can we learn from Squid Game about the types of shows that are successful?

What Good is a Regression?

To answer these questions, we spent last week building up to the idea of the conditional expectation function (CEF) of Y:

$$\mathbb{E}[Y|X] = f(X)$$

- The CEF gives the expected value of Y for a given value of $X = x$

What Good is a Regression?

To answer these questions, we spent last week building up to the idea of the conditional expectation function (CEF) of Y:

$$\mathbb{E}[Y|X] = f(X)$$

- The CEF gives the expected value of Y for a given value of $X = x$
- Why do we care? Because many of those previous questions are answered via comparison:
  - Bivariate: Comparing values of Y for different values of X
  - Multivariate: Comparing values of Y for different combinations of features

What Good is a Regression?

To answer these questions, we spent last week building up to the idea of the conditional expectation function (CEF) of Y:

$$\mathbb{E}[Y|X] = f(X)$$

- The CEF gives the expected value of Y for a given value of $X = x$
- Why do we care? Because many of those previous questions are answered via comparison:
  - Bivariate: Comparing values of Y for different values of X
  - Multivariate: Comparing values of Y for different combinations of features
- We can use regression (bivariate or multivariate) to make these comparisons

What Good is a Regression?

Recall that we can use regression in at least three ways:

1. As a **causal device**:
   - ➜ Estimate the effect of a feature on the outcome
   - ➜ Comparison: if X changes, does this *change* Y?
   - ➜ E.g. 'Did Squid Game increase Netflix viewership?'

What Good is a Regression?

Recall that we can use regression in at least three ways:

1. As a **causal device**:
   - ➜ Estimate the effect of a feature on the outcome
   - ➜ Comparison: if X changes, does this *change* Y?
   - ➜ E.g. 'Did Squid Game increase Netflix viewership?'

2. As a **descriptive device**:
   - ➜ Summarize co-occurence of two variables, irrespective DGP
   - ➜ Comparison: for different levels of X, is Y usually different?
   - ➜ E.g. 'Which viewers most liked Squid Game?'

What Good is a Regression?

Recall that we can use regression in at least three ways:

1. As a **causal device**:
   - ➜ Estimate the effect of a feature on the outcome
   - ➜ Comparison: if X changes, does this *change* Y?
   - ➜ E.g. 'Did Squid Game increase Netflix viewership?'

2. As a **descriptive device**:
   - ➜ Summarize co-occurence of two variables, irrespective DGP
   - ➜ Comparison: for different levels of X, is Y usually different?
   - ➜ E.g. 'Which viewers most liked Squid Game?'

3. As a **predictive device**:
   - ➜ Fill in 'missing values' (unseen realizations) of Y using X
   - ➜ Comparison: if we saw a new value of X, what would we expect Y to be?
   - ➜ E.g. 'Who should we advertise or push Squid Game to?'

Goals for Today

1. When can we feel comfortable making claims of type 1?
   - ➜ The DGP of Y
   - ➜ The assignment mechanism for D

Goals for Today

1. When can we feel comfortable making claims of type 1?
   - ➜ The DGP of Y
   - ➜ The assignment mechanism for D

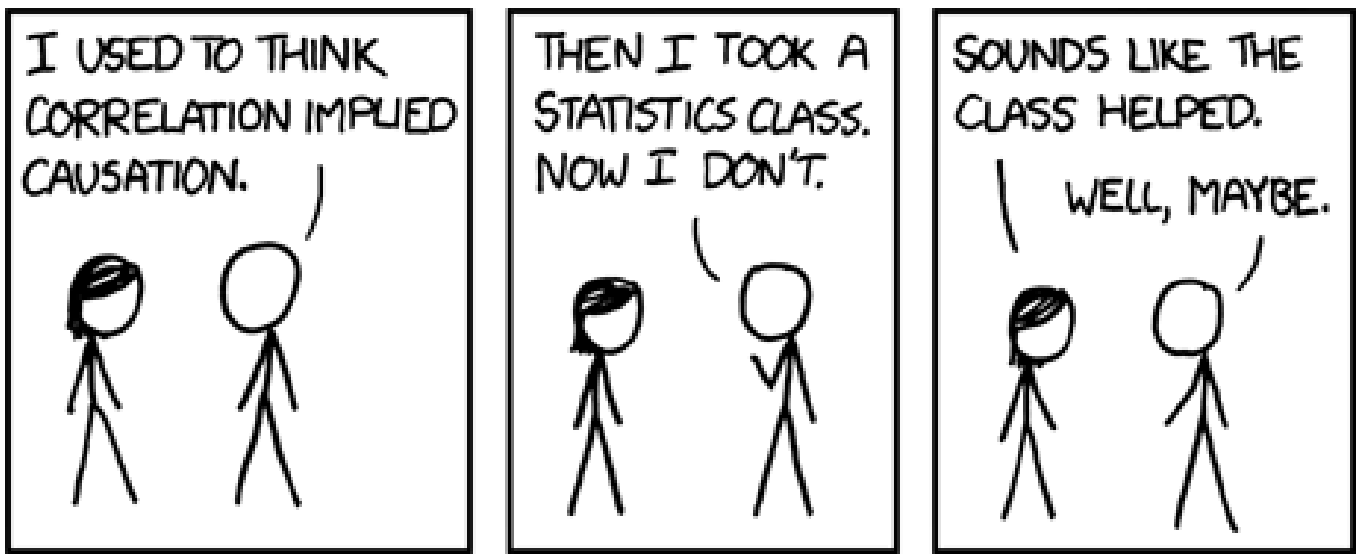
2. How can we make those claims using data?
   - ➜ Matching
   - ➜ Regression (woo!)

Goals for Today

1. When can we feel comfortable making claims of type 1?
   - ➜ The DGP of Y
   - ➜ The assignment mechanism for D

2. How can we make those claims using data?
   - ➜ Matching
   - ➜ Regression (woo!)

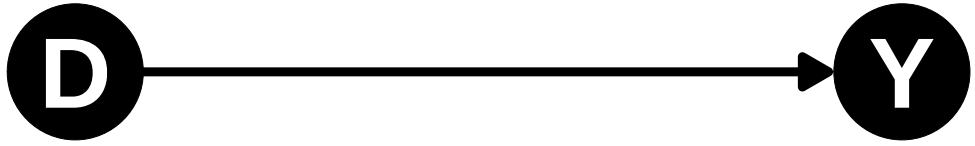
3.



Source: Randall Monroe, https://xkcd.com/552/

Theories of Causation

Cause and Effect



D: Our 'causal feature/variable,' often called a 'treatment'

Y: Our 'outcome variable', or 'response variable'

Cause and Effect

Before we can get to the good stuff, we have to do a little theory:

1. Define causes and effects in abstract terms

2. Build two formal representations of causation:

   - ➜ Potential outcomes
   - ➜ Graphical models

Cause and Effect

Causes and their effects have two properties: they are **successive** and can be reasoned about in **counterfactual** terms:

Cause and Effect

Causes and their effects have two properties: they are **successive** and can be reasoned about in **counterfactual** terms:

*[. . .] We may define a cause to be an object **followed by** another, [. . .] where, if the first object **had not been**, the second never had existed.* – Hume, 1748

Cause and Effect

Causes and their effects have two properties: they are **successive** and can be reasoned about in **counterfactual** terms:

*[. . .] We may define a cause to be an object **followed by** another, [. . .] where, if the first object **had not been**, the second never had existed.* – Hume, 1748

*[. . .] would not have died **if he had not** eaten of it, people would be apt to say that eating of that dish was the cause of his death.* – Mill, 1843

Cause and Effect

Causes and their effects have two properties: they are **successive** and can be reasoned about in **counterfactual** terms:

*[. . .] We may define a cause to be an object **followed by** another, [. . .] where, if the first object **had not been**, the second never had existed.* – Hume, 1748

*[. . .] would not have died **if he had not** eaten of it, people would be apt to say that eating of that dish was the cause of his death.* – Mill, 1843

One important implication is that causal variables must be **manipulable**:

Cause and Effect

Causes and their effects have two properties: they are **successive** and can be reasoned about in **counterfactual** terms:

*[. . .] We may define a cause to be an object **followed by** another, [. . .] where, if the first object **had not been**, the second never had existed.* – Hume, 1748

*[. . .] would not have died **if he had not** eaten of it, people would be apt to say that eating of that dish was the cause of his death.* – Mill, 1843

One important implication is that causal variables must be **manipulable**:

*No **causation** without **manipulation.*** – Holland, 1986

Good Causal Questions

Manipulability means we must think very carefully about causal questions. . .

1. (Largely) immutable characteristics:

- ➜ Customers' sex assigned at birth $\rightarrow$ consumer preferences
- ➜ Race and ethnicity $\rightarrow$ employment outcomes
- ➜ Country of origin $\rightarrow$ ideology

Good Causal Questions

Manipulability means we must think very carefully about causal questions. . .

1. (Largely) immutable characteristics:

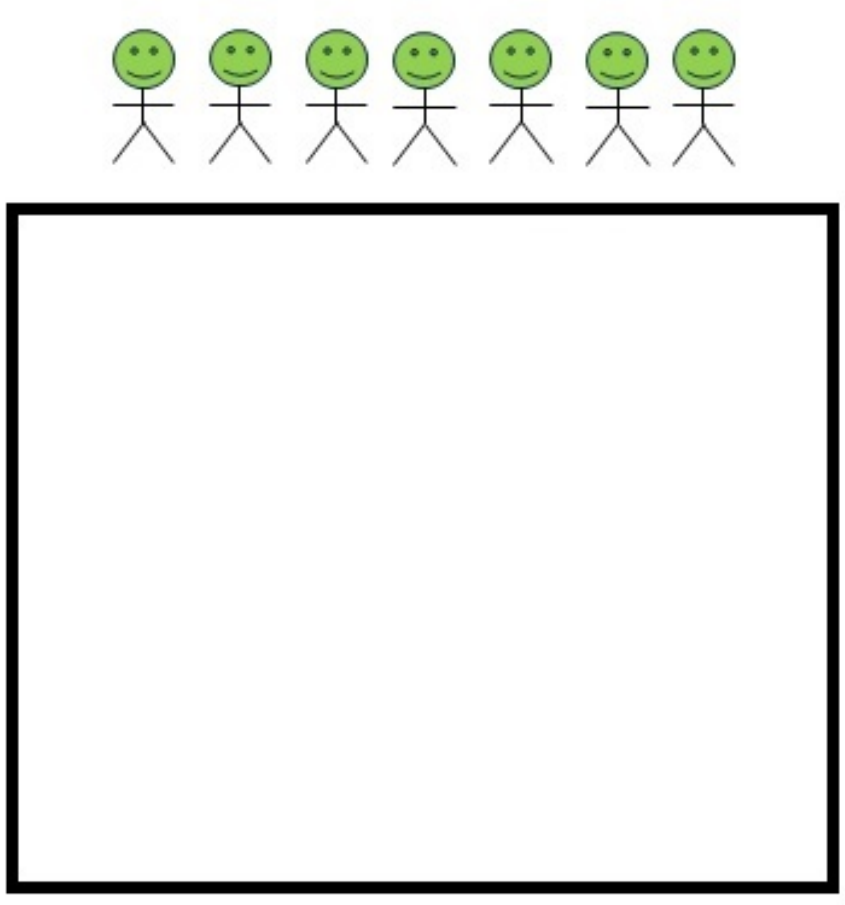
➜ Customers' sex assigned at birth $\rightarrow$ consumer preferences
➜ Race and ethnicity $\rightarrow$ employment outcomes
➜ Country of origin $\rightarrow$ ideology

2. Non-successive chains:

➜ Monthly expenditure $\rightarrow$ monthly savings
➜ Platform decision made in 2012 $\rightarrow$ customer behaviour today

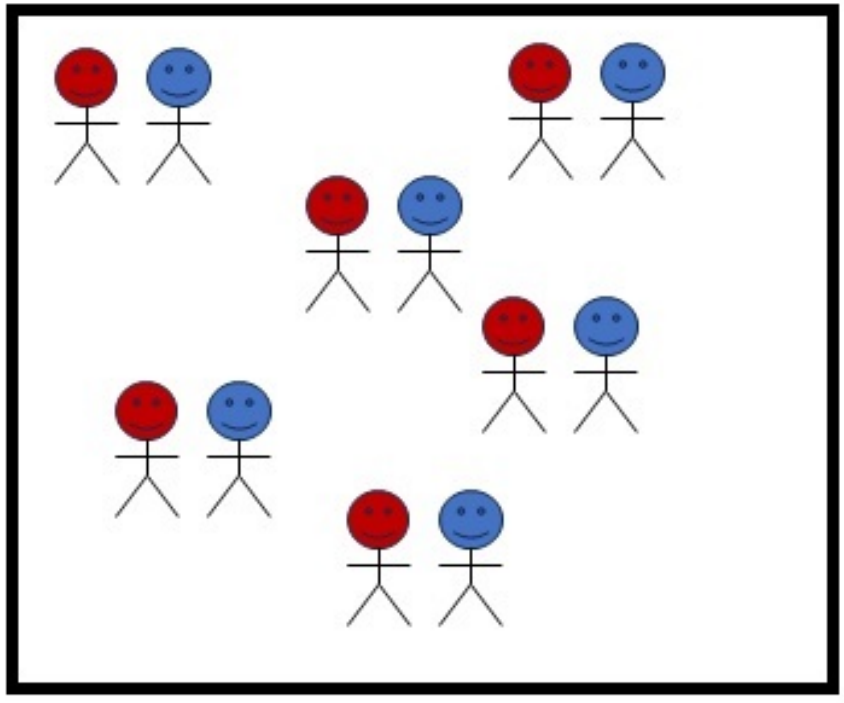
Potential Outcomes

We have six participants in a study with a binary treatment:



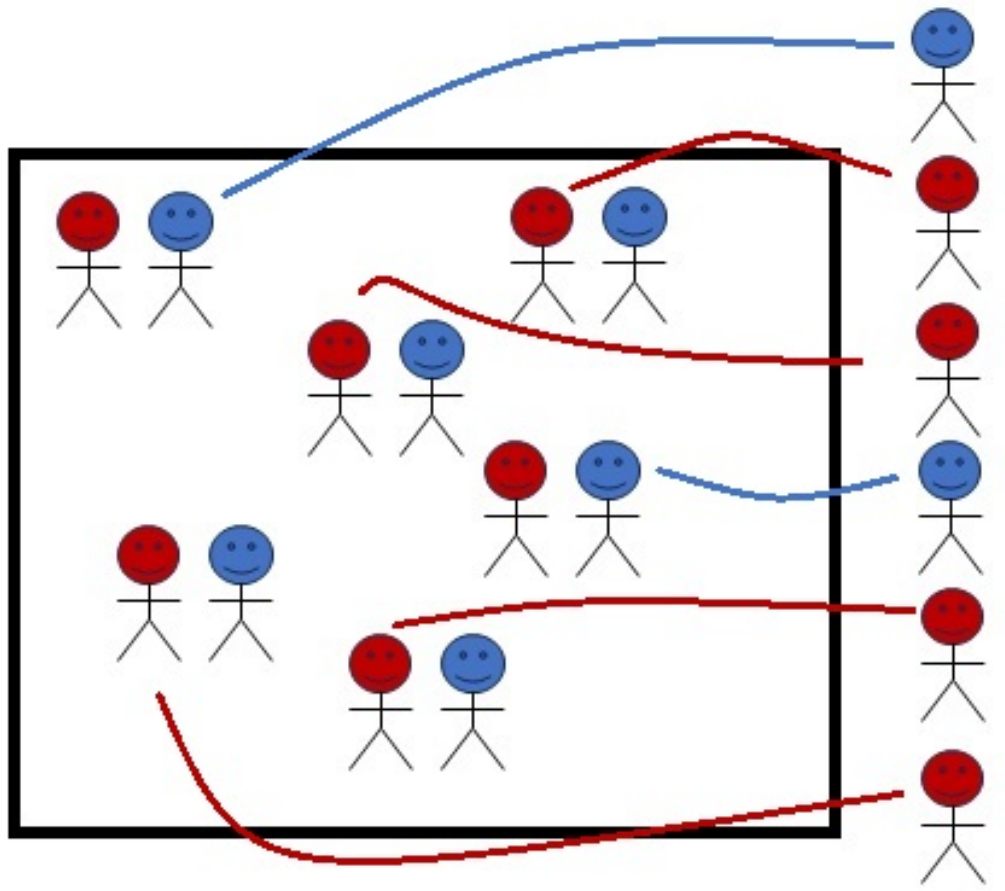
Potential Outcomes

Each has **two** potential outcomes:



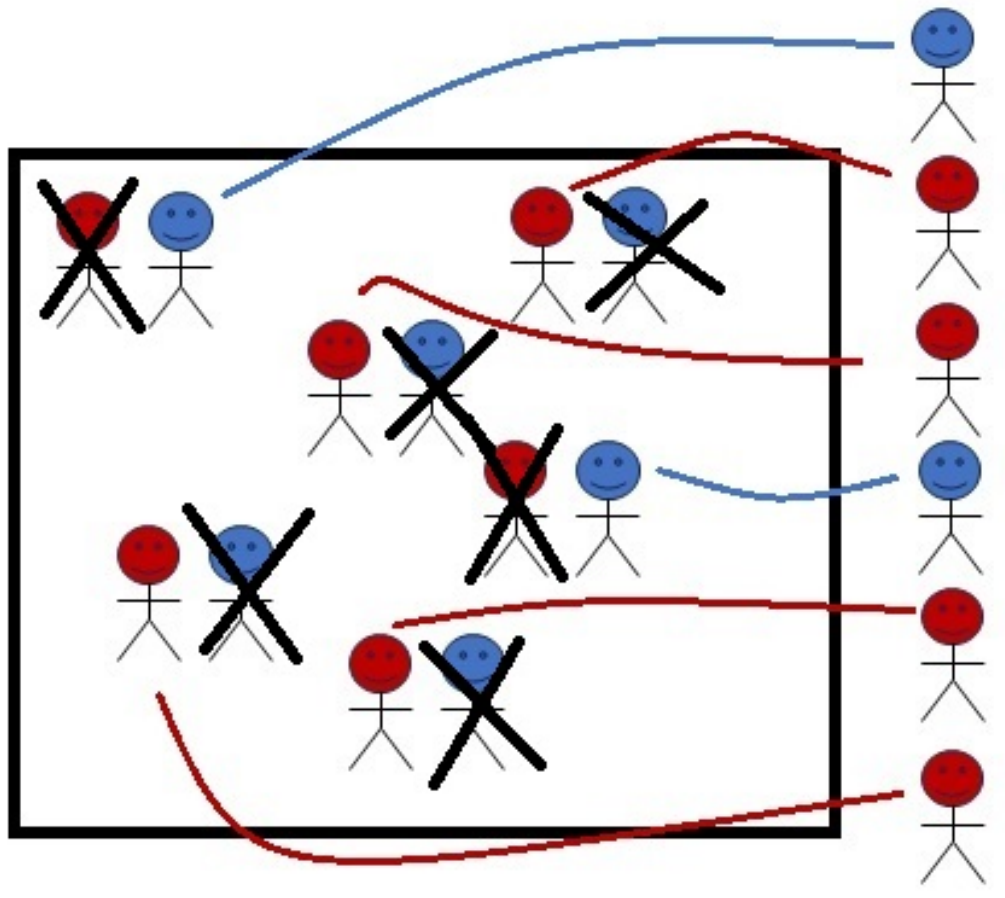
Potential Outcomes

Only **one** is realised, based on treatment status:



Potential Outcomes

The **Fundamental Problem of Causal Inference** (FPCI):



Potential Outcomes

Let's do this more formally, assuming some treatment D on an outcome Y.

Potential Outcomes

Let's do this more formally, assuming some treatment D on an outcome Y.

For every individual $i = 1, \ldots, N$, we say they have **potential outcomes** written as $Y_i(D_i = d)$

Potential Outcomes

Let's do this more formally, assuming some treatment D on an outcome Y.

For every individual $i = 1, \ldots, N$, we say they have **potential outcomes** written as $Y_i(D_i = d)$

For a binary treatment ($D \in \{0, 1\}$), two potential outcomes:

- $Y_i(0)$: The value Y would take if they did not receive treatment ($D_i = 0$)
- $Y_i(1)$: The value Y would take if they did receive treatment ($D_i = 1$)

These are mutually exclusive **counterfactual** quantities: Only one can ever be realized.

Potential Outcomes

Let's go back to the toy example, but this time in R.

We have six students, each with two potential outcomes:

```r
students <- data.frame(
  student = c("Radha", "Pam", "Konstantinos",
              "Joy", "Shawn", "Brooke"),
  Headache_0 = c(9, 4, 8, 3, 10, 3),
  Headache_1 = c(7, 1, 9, 5, 7, 3)
)

students
```

```
       student Headache_0 Headache_1
1        Radha          9          7
2          Pam          4          1
3 Konstantinos          8          9
4          Joy          3          5
5        Shawn         10          7
6       Brooke          3          3
```

Potential Outcomes

So, let's distribute or 'assign' a **treatment**:

```
students <- students %>%
  mutate(Medicine = sample(c(0, 1), nrow(students), replace = TRUE))

students
```

```
       student Headache_0 Headache_1 Medicine
1        Radha          9          7        1
2          Pam          4          1        1
3 Konstantinos          8          9        1
4          Joy          3          5        0
5        Shawn         10          7        1
6       Brooke          3          3        1
```

Potential Outcomes

The **realized outcomes** are a function of potential outcomes and treatment:

```
students <- students %>%
  mutate(Headache = case_when(
            Medicine == 0 ~ Headache_0,
            Medicine == 1 ~ Headache_1,
            TRUE ~ NA_real_
          )
  )

students
```

```
        student Headache_0 Headache_1 Medicine Headache
1         Radha          9          7        1        7
2           Pam          4          1        1        1
3 Konstantinos          8          9        1        9
4           Joy          3          5        0        3
5         Shawn         10          7        1        7
6        Brooke          3          3        1        3
```

Potential Outcomes to Realized Outcomes

Formally, the realized or observed outcomes are:

$$Y_i = Y_i(1) \times D_i + Y_i(0) \times (1 - D_i)$$

Read:

- When treatment is 1, we get back $Y_i(1)$
- When treatment is 0, we get back $Y_i(0)$.

Potential Outcomes: Estimands

Remember that in general we want to be precise about the **estimands** we are targeting.

Potential Outcomes: Estimands

Remember that in general we want to be precise about the **estimands** we are targeting.

Let's define the **Individual Treatment Effect (ITE)**:

$$ITE_i = Y_i(1) - Y_i(0)$$

Read: For individual i, the effect of of treatment is the difference in that individual's potential outcomes.

Potential Outcomes: Estimands

Remember that in general we want to be precise about the **estimands** we are targeting.

Let's define the **Individual Treatment Effect (ITE)**:

$$ITE_i = Y_i(1) - Y_i(0)$$

Read: For individual i, the effect of of treatment is the difference in that individual's potential outcomes.

Intuition check: When would $Y_i = Y_i(1) = Y_i(0)$?

Potential Outcomes: Estimands

Remember that in general we want to be precise about the **estimands** we are targeting.

Let's define the **Individual Treatment Effect (ITE)**:

$$ITE_i = Y_i(1) - Y_i(0)$$

Read: For individual i, the effect of of treatment is the difference in that individual's potential outcomes.

Intuition check: When would $Y_i = Y_i(1) = Y_i(0)$?

When i does not respond to treatment for, in other words: $ITE_i = 0$.

Potential Outcomes: Estimands

Individual-level effects are very hard to target. (Why?)

Instead let's focus on two group-level estimands:

Potential Outcomes: Estimands

Individual-level effects are very hard to target. (Why?)

Instead let's focus on two group-level estimands:

Average Treatment Effect (ATE):

$$ATE = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$$

Average Treatment Effect on the Treated (ATT):

$$ATT = \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1]$$

Potential Outcomes: Estimands

Let's go back to our toy example to see this more clearly:

```
students %>%
  summarise(ATE = mean(Headache_1) - mean(Headache_0),
            ATT = mean(Headache_1[Medicine == 1]) -
              mean(Headache_0[Medicine == 1]),
            ATE_hat = mean(Headache[Medicine == 1]) -
              mean(Headache[Medicine == 0]))
```

```
          ATE  ATT ATE_hat
1 -0.8333333 -1.4     2.4
```

Graphical Models of Causation

An alternative (but $\sim$ equivalent) way to represent causal relationships is with a **graphical model**.

We call these **Directed Acyclic Graphs** (DAGs).

Graphical Models of Causation

An alternative (but $\sim$ equivalent) way to represent causal relationships is with a **graphical model**.

We call these **Directed Acyclic Graphs** (DAGs).

For any causal question, we can write down our beliefs about:

1. The data generating process for Y
2. The data generating process for D

Graphical Models of Causation

For our purposes, a graphical causal model has four features:

- **Nodes**: represent features or outcomes (e.g., X, Y)
- **Edges**: represent causal relationships (e.g., $X \rightarrow Y$)
  - Note, the *absence* of edges implies an *absence* of a causal relationship
- **Directed**: Pairs can be said to have a tail and a head (and in chains, ancestors and descendants)
- **Acyclical**: Nodes cannot terminate in themselves (e.g., $X \rightarrow Y \rightarrow X$)

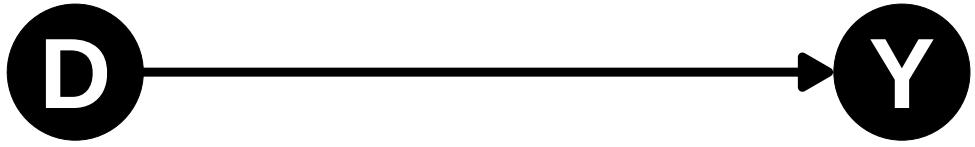
Graphical Models of Causation

For our purposes, a graphical causal model has four features:

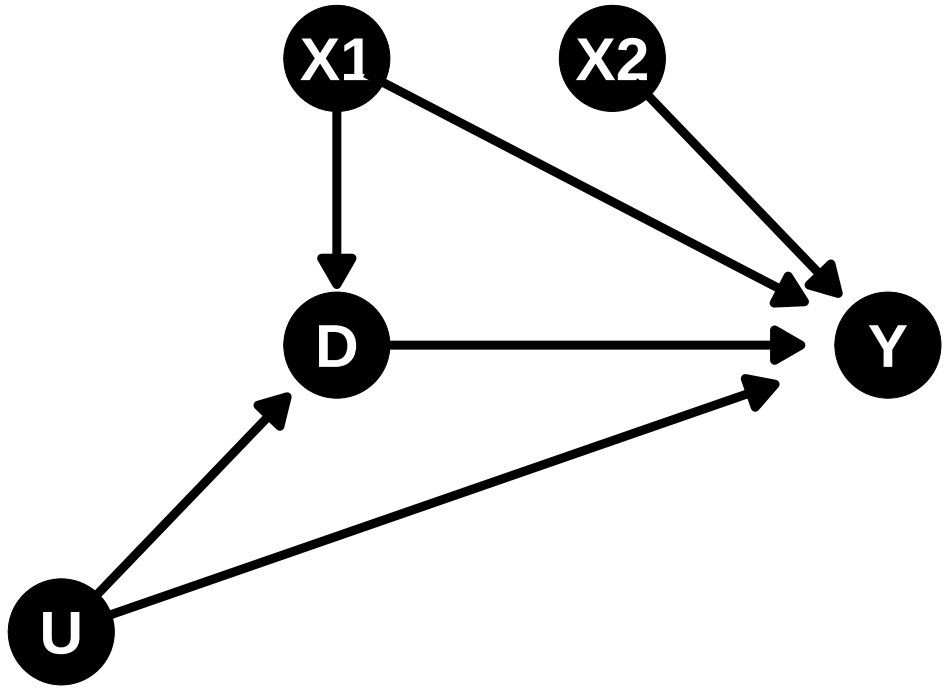
- **Nodes**: represent features or outcomes (e.g., X, Y)
- **Edges**: represent causal relationships (e.g., $X \rightarrow Y$)
  - Note, the *absence* of edges implies an *absence* of a causal relationship
- **Directed**: Pairs can be said to have a tail and a head (and in chains, ancestors and descendants)
- **Acyclical**: Nodes cannot terminate in themselves (e.g., $X \rightarrow Y \rightarrow X$)

One key thing to remember: A graphical model is a *theory of the DGPs of Y and D* that you *assume*.

Graphical Models of Causation



Graphical Models of Causation



The Identification Problem

The Identification Problem for Causal Inference

Identification:

In statistics, an **estimand** (parameter) is **identified** if its value can, asymptotically, be uniquely **mapped to** observed data and unidentified otherwise.

If an estimand is not identified, we can say there are **alternative explanations** (mappings) connecting the observed data and the estimand.

The Identification Problem for Causal Inference

Identification:

In statistics, an **estimand** (parameter) is **identified** if its value can, asymptotically, be uniquely **mapped to** observed data and unidentified otherwise.

If an estimand is not identified, we can say there are **alternative explanations** (mappings) connecting the observed data and the estimand.

For causal questions, estimands are typically population causal effects but the **FPCI** tells us that at least half of the potential outcomes are always missing.

The Identification Problem for Causal Inference

Identification:

In statistics, an **estimand** (parameter) is **identified** if its value can, asymptotically, be uniquely **mapped to** observed data and unidentified otherwise.

If an estimand is not identified, we can say there are **alternative explanations** (mappings) connecting the observed data and the estimand.

For causal questions, estimands are typically population causal effects but the **FPCI** tells us that at least half of the potential outcomes are always missing.

Let's see this problem formally.

Selection Bias

The naïve difference of observed means in the treatment groups:

$$\underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{Observed difference in average outcome measures}} = \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$$

Selection Bias

The naïve difference of observed means in the treatment groups:

$$\underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{Observed difference in average outcome measures}} = \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$$

$$= \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

Read: The same observed mean difference could be due to **different combinations** of the ATT (estimand!) and selection bias terms.

Selection Bias

The naïve difference of observed means in the treatment groups:

$$\underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{Observed difference in average outcome measures}} = \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$$

$$= \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

Read: The same observed mean difference could be due to **different combinations** of the ATT (estimand!) and selection bias terms.

Thus ATT is **unidentified** from the naïve observed mean difference: it is not uniquely mapped from the observed data.

Selection Bias

The naïve difference of observed means in the treatment groups:

$$\underbrace{\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]}_{\text{Observed difference in average outcome measures}} = \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$$

$$= \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

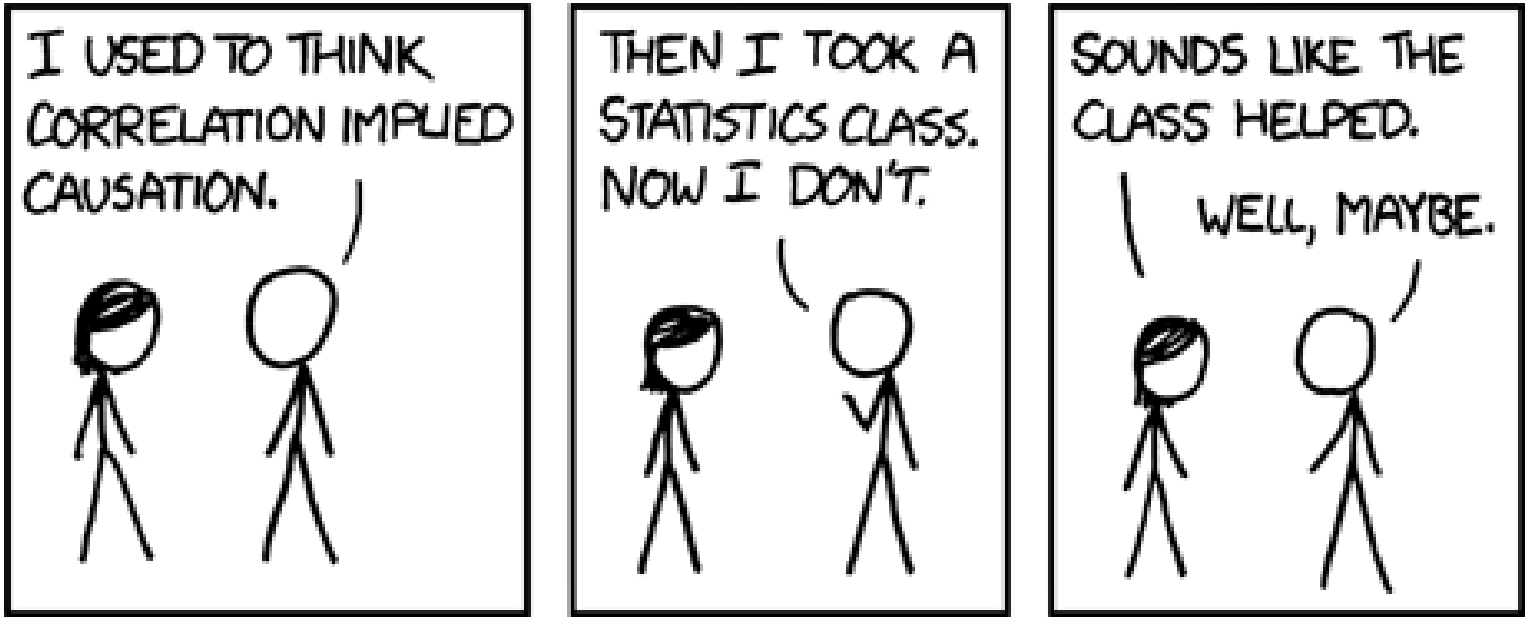
Read: The same observed mean difference could be due to **different combinations** of the ATT (estimand!) and selection bias terms.

Thus ATT is **unidentified** from the naïve observed mean difference: it is not uniquely mapped from the observed data.

Correlation (association, observed difference) is not necessarily causation.

Selection Bias

Haha!



Source: Randall Monroe, https://xkcd.com/552/

Selection Bias

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}}$$
$$+ \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

$\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$ is called **selection bias** because, if it is not zero, treatment and control groups are systematically different in Y_{0i}.

If non-zero, we might say the causal effect of D on Y is **confounded**.

Selection Bias

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}}$$
$$+ \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

$\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]$ is called **selection bias** because, if it is not zero, treatment and control groups are systematically different in Y_{0i}.

If non-zero, we might say the causal effect of D on Y is **confounded**.

Canonical example – those who are more risk averse will be more likely to wear a seatbelt:

$$\mathbb{E}[Y_0 \mid D = 1] - \mathbb{E}[Y_0 \mid D = 0] > 0$$

The Identification Problem Redux

Identification Strategy:

A combination of **data** and **assumptions** which allows us to **identify** a causal estimand by estimating ("filling in") the missing potential outcomes (usually at a group level) in expectation.

The Identification Problem Redux

Identification Strategy:

A combination of **data** and **assumptions** which allows us to **identify** a causal estimand by estimating ("filling in") the missing potential outcomes (usually at a group level) in expectation.

Today:

- ➜ Randomization
- ➜ Selection on Observables (SOO)

Tomorrow:

- ➜ Instrumental Variables (IV)
- ➜ Regression Discontinuity (RD)
- ➜ Difference-in-Differences (DiD)

Randomization

Randomization: Setup

As before, our setting:

- Y: Outcome variable
- $D \in \{0, 1\}$: Treatment variable (binary)
- $Y(0)$: Potential outcome if $D = 0$
- $Y(1)$: Potential outcome if $D = 1$

Assume that D is assigned by an understood random **assignment mechanism**

Randomization: Setup

As before, our setting:

- Y: Outcome variable
- $D \in \{0, 1\}$: Treatment variable (binary)
- $Y(0)$: Potential outcome if $D = 0$
- $Y(1)$: Potential outcome if $D = 1$

Assume that D is assigned by an understood random **assignment mechanism**

This is called a 'randomized experiment', 'randomized controlled trial' (RCT), 'A/B test'.

Randomization: Setup

As before, our setting:

- Y: Outcome variable
- $D \in \{0, 1\}$: Treatment variable (binary)
- $Y(0)$: Potential outcome if $D = 0$
- $Y(1)$: Potential outcome if $D = 1$

Assume that D is assigned by an understood random **assignment mechanism**

This is called a 'randomized experiment', 'randomized controlled trial' (RCT), 'A/B test'.

There are many ways to do this (simple, complete, clustered, etc.).

Randomization: Independence

Assumption: $D \perp \{Y(0), Y(1)\}$

Read: Treatment status (D) is independent of potential outcomes.

We will call this assumption **independence** (aka ignorability, randomization)

Intuition: By randomizing treatment, we have severed any relationship between **treatment** and **potential outcomes**.

Check: Does the independence assumption imply $D \perp Y$?

Randomization: Independence

Assumption: $D \perp \{Y(0), Y(1)\}$

Read: Treatment status (D) is independent of potential outcomes.

We will call this assumption **independence** (aka ignorability, randomization)

Intuition: By randomizing treatment, we have severed any relationship between **treatment** and **potential outcomes**.

Check: Does the independence assumption imply $D \perp Y$? No!

Randomization: Identification (POs)

How does independence help us? Recall selection bias:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}}$$
$$+ \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

Randomization: Identification (POs)

How does independence help us? Recall selection bias:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \underbrace{\mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]}_{\text{ATT}} + \underbrace{\mathbb{E}[Y_{0i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 0]}_{\text{Selection bias}}$$

If $D \perp \{Y(0), Y(1)\}$, then:

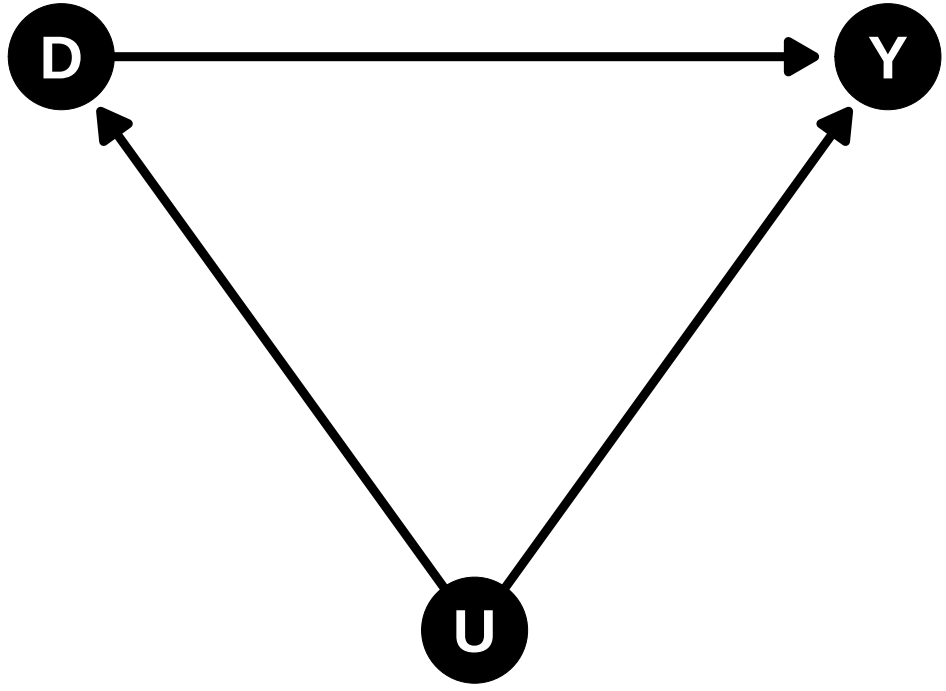
$$\mathbb{E}[Y_{0i} \mid D_i = 1] = \mathbb{E}[Y_{0i} \mid D_i = 0]$$

Thus, the selection bias term is zero and:

$$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] = \mathbb{E}[Y_{1i} \mid D_i = 1] - \mathbb{E}[Y_{0i} \mid D_i = 1]$$

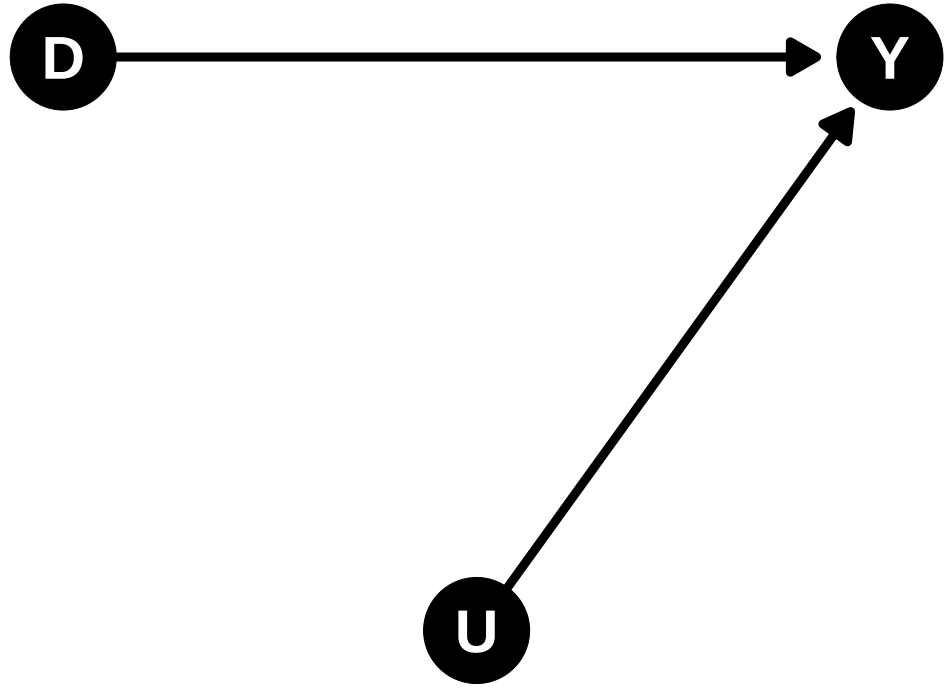
QED

Randomization: Identification (DAGs)



Here U is a canonical **confounder** – it both **selects** individuals into treatment and **affects** the outcome.

Randomization: Identification (DAGs)



Under randomization, the $U \rightarrow D$ is severed – there is no longer confounding.

Randomization: Estimation

In a randomized experiment, assuming independence, we have proven that:

The observed **difference in means** (D-i-M) is an unbiased estimator of the ATT.

Randomization: Estimation

In a randomized experiment, assuming independence, we have proven that:

The observed **difference in means** (D-i-M) is an unbiased estimator of the ATT.

It turns out that because D is randomly assigned, it is *also* an unbiased estimator of the ATE:

$$\begin{aligned}
\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] &= \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0] \\
&= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] \qquad \text{(by independence)} \\
&= \mathbb{E}[Y_i(1) - Y_i(0)]
\end{aligned}$$

So, in a randomized experiment, you can just calculate the group D-i-M, and use a t-test.

Randomization: Estimation

Let's do just that on a slightly larger dataset than our toy example:

```
t.test(Headache ~ Medicine, data = students_analysis,
       var.equal = FALSE)
```

```
    Welch Two Sample t-test

data:  Headache by Medicine
t = -17.658, df = 918.16, p-value < 2.2e-16
alternative hypothesis: true difference in means between group 0 and group 1 is
95 percent confidence interval:
 -3.788941 -3.030952
sample estimates:
mean in group 0 mean in group 1
       5.614484        9.024431
```

Randomization: Estimation

An alternative (but equivalent) estimator is linear regression:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$$

Under independence, $\hat{\beta}_1$ identifies the ATT and ATE.

Randomization: Estimation

An alternative (but equivalent) estimator is linear regression:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$$

Under independence, $\hat{\beta}_1$ identifies the ATT and ATE.

This is not surprising as bivariate regression with a binary explanatory feature *is* the difference-in-means.

Remember: Regression is a tool for comparison!

Randomization: Estimation

An alternative (but equivalent) estimator is linear regression:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$$

Under independence, $\hat{\beta}_1$ identifies the ATT and ATE.

This is not surprising as bivariate regression with a binary explanatory feature *is* the difference-in-means.

Remember: Regression is a tool for comparison!

Intuition check: What does $\hat{\beta}_0$ identify?

Randomization: Estimation

An alternative (but equivalent) estimator is linear regression:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i$$

Under independence, $\hat{\beta}_1$ identifies the ATT and ATE.

This is not surprising as bivariate regression with a binary explanatory feature *is* the difference-in-means.

Remember: Regression is a tool for comparison!

Intuition check: What does $\hat{\beta}_0$ identify? $\mathbb{E}[Y_i(0)]$

Randomization: Asymptotic Inference

In a randomized experiment, we can use the same results we saw last week to do inference on our estimates of the ATE and ATT.

If using regression, use heteroskedasticity-robust (or clustered, if needed) standard errors.

Putting it all together:

```
lm_robust(Headache ~ Medicine, data = students_analysis) %>%
  summary()
```

```
Call:
lm_robust(formula = Headache ~ Medicine, data = students_analysis)

Standard error type:  HC2

Coefficients:
            Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
(Intercept)    5.614     0.1152   48.76 2.637e-266    5.389    5.840 998
Medicine       3.410     0.1931   17.66  6.222e-61    3.031    3.789 998

Multiple R-squared:  0.2387 ,   Adjusted R-squared:  0.2379
```

Selection on Observables

SOO: Setup

In a randomized experiment **we** control the assignment mechanism. Often that is not the case.

As before, our setting:

- Y: Outcome variable
- $D \in \{0, 1\}$: Treatment variable (binary)
- $Y(0)$: Potential outcome if $D = 0$
- $Y(1)$: Potential outcome if $D = 1$
- X: A (set of) observable pre-treatment covariate(s)

D is not randomized, so we have to rely on **observed variation in** D, and make some assumptions about it.

SOO: Setup

In a randomized experiment **we** control the assignment mechanism. Often that is not the case.

As before, our setting:

- Y: Outcome variable
- $D \in \{0, 1\}$: Treatment variable (binary)
- $Y(0)$: Potential outcome if $D = 0$
- $Y(1)$: Potential outcome if $D = 1$
- X: A (set of) observable pre-treatment covariate(s)

D is not randomized, so we have to rely on **observed variation in** D, and make some assumptions about it.

This is called **Selection on Observables** (SOO): We believe the selection process is a function of observed features.

SOO: Conditional Independence

Assumption 1: $D \perp \{Y(0), Y(1)\} \mid X \quad$ for any $\quad x \in \mathcal{X}$

Read: Treatment status (D) is independent of potential outcomes, conditional on X.

We will call this assumption **conditional independence** (aka conditional ignorability).

SOO: Conditional Independence

Assumption 1: $D \perp \{Y(0), Y(1)\} \mid X \quad$ for any $\quad x \in \mathcal{X}$

Read: Treatment status (D) is independent of potential outcomes, conditional on X.

We will call this assumption **conditional independence** (aka conditional ignorability).

Intuition: Proposes that within each level of X, there is an 'experiment' such that $D \perp \{Y(0), Y(1)\}$.

SOO: Common Support

Assumption 2: $0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \text{for any} \quad x \in \mathcal{X}$

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

Intuition: Proposes that each experiment is 'meaningful'

SOO: Common Support

Assumption 2: $0 < \Pr(D_i = 1 \mid X_i = x) < 1 \quad \text{for any} \quad x \in \mathcal{X}$

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

Intuition: Proposes that each experiment is 'meaningful'

We will omit the proof, but here is the intuition:

- ➜ The effect of D on Y for each value of X is called a **Conditional Average Treatment Effect (CATE)**
- ➜ Each CATE is identified (by conditional independence)
- ➜ The ATE (or ATT) is just a weighted average of all CATEs (by common support)
- ➜ The ATE is identified as a weighted average of CATEs, where weights $\rightarrow$ $P(X_i = x)$

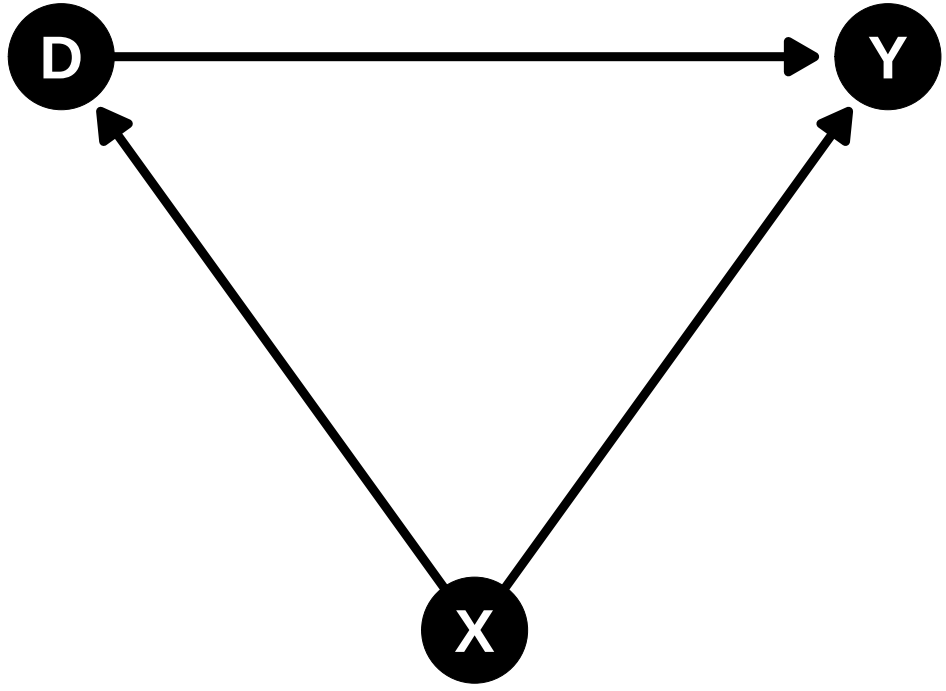
SOO: Identification (DAGs)

We can use DAGs to learn about identification conditions in SOO settings.

We want to choose a **conditioning set** (X) that **blocks all back-door paths** between D and Y:

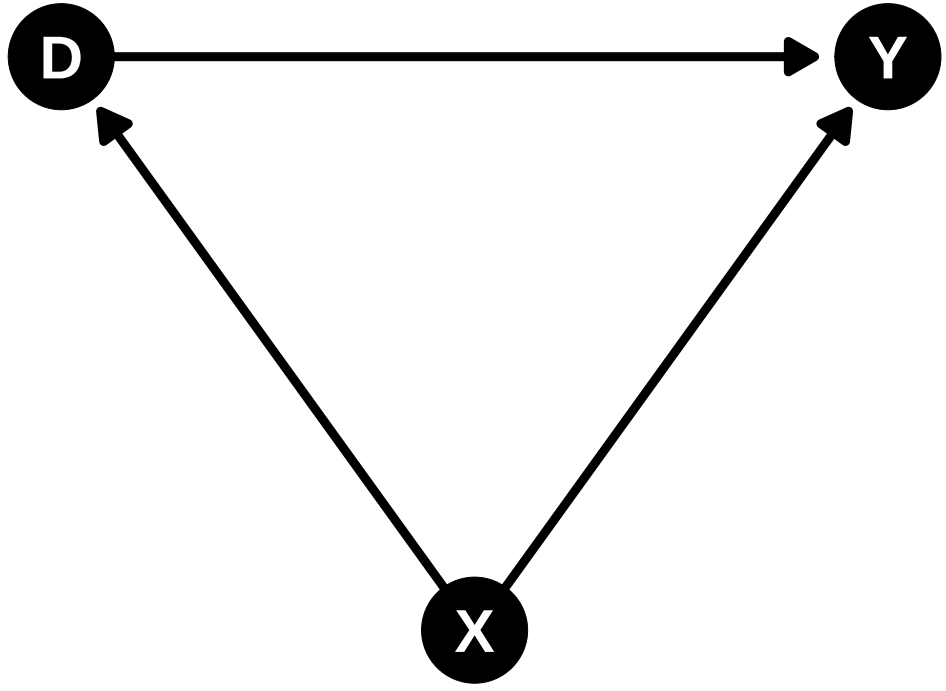
- ➜ Back-door path: A path that starts with an arrow into D and ends with an arrow into Y, includes no descendants of D, and does not have a collider.
- ➜ Collider: A node with two arrows into it, e.g., $D \rightarrow U \leftarrow Y$. This blocks a path.

SOO: Identification (DAGs)



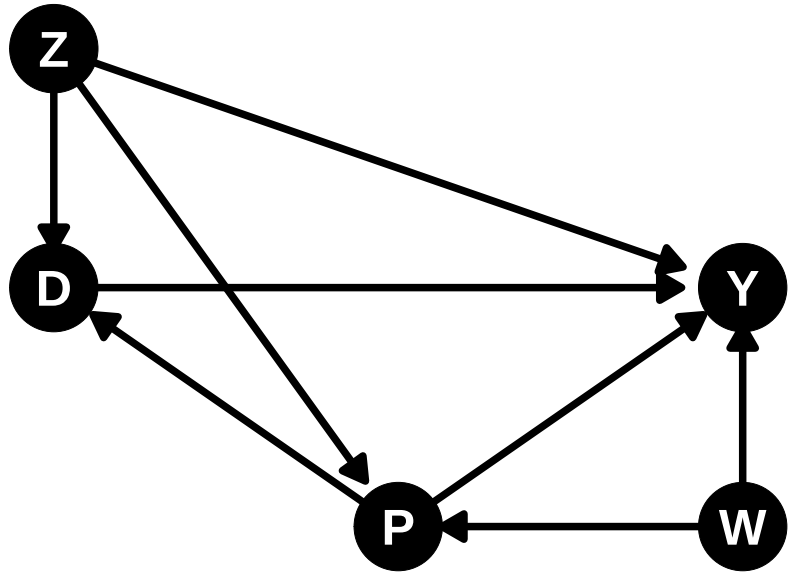
Is $\{X\}$ is a sufficient conditioning set to identify the effect of D on Y.

SOO: Identification (DAGs)



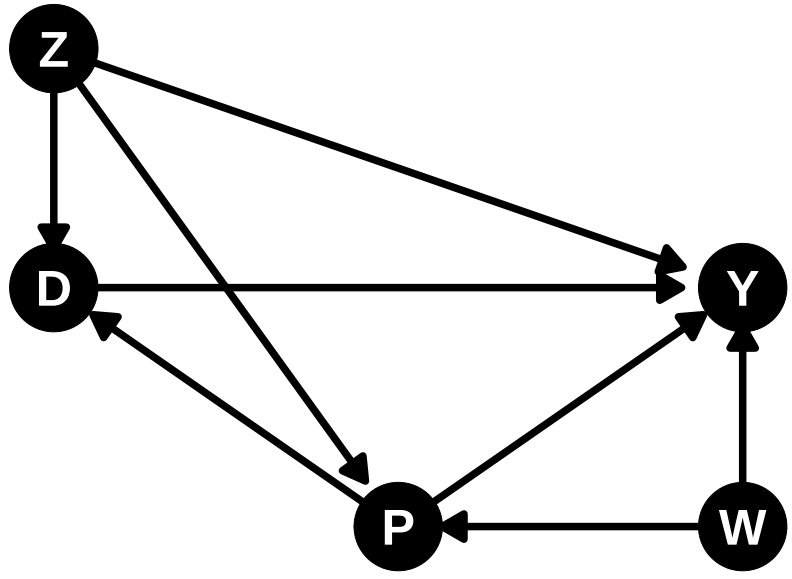
Is $\{X\}$ is a sufficient conditioning set to identify the effect of D on Y. Yes!

SOO: Identification (DAGs)



Does $\{Z\}$ identify the effect of D on Y?

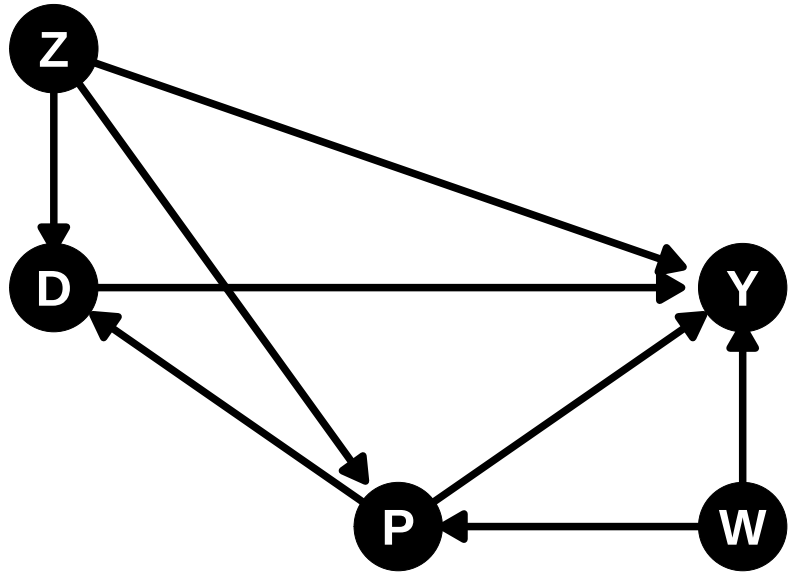
SOO: Identification (DAGs)



Does $\{Z\}$ identify the effect of D on Y? No!

Does $\{P\}$ identify the effect of D on Y?

SOO: Identification (DAGs)

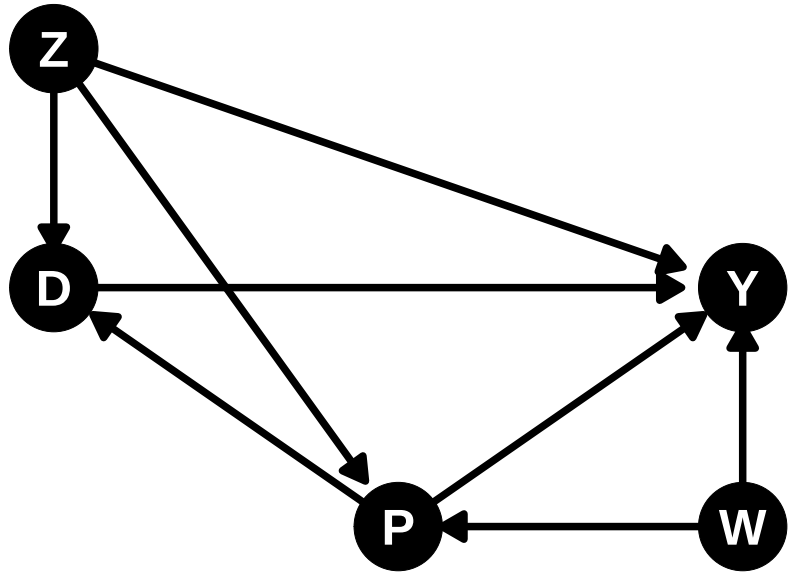


Does $\{Z\}$ identify the effect of D on Y? No!

Does $\{P\}$ identify the effect of D on Y? No!

Does $\{Z, P\}$ identify the effect of D on Y?

SOO: Identification (DAGs)



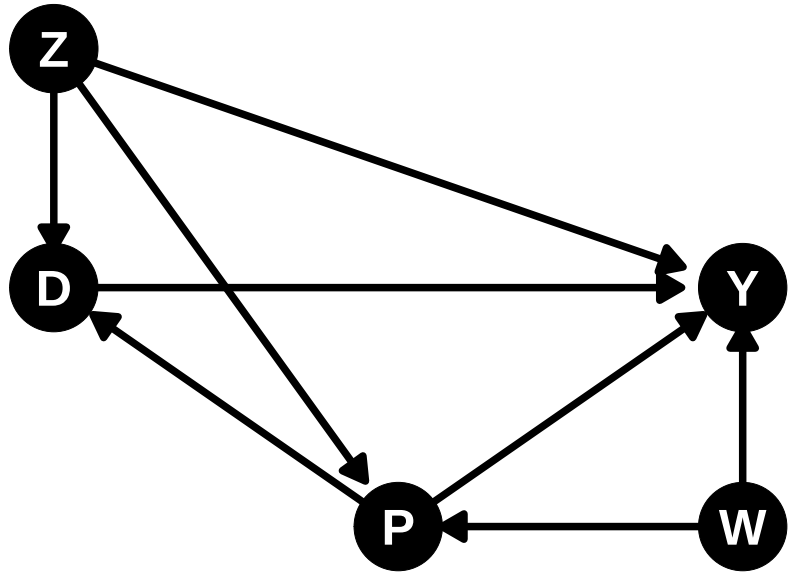
Does $\{Z\}$ identify the effect of D on Y? No!

Does $\{P\}$ identify the effect of D on Y? No!

Does $\{Z, P\}$ identify the effect of D on Y? Yes!

Does $\{Z, W\}$ identify the effect of D on Y?

SOO: Identification (DAGs)



Does $\{Z\}$ identify the effect of D on Y? No!

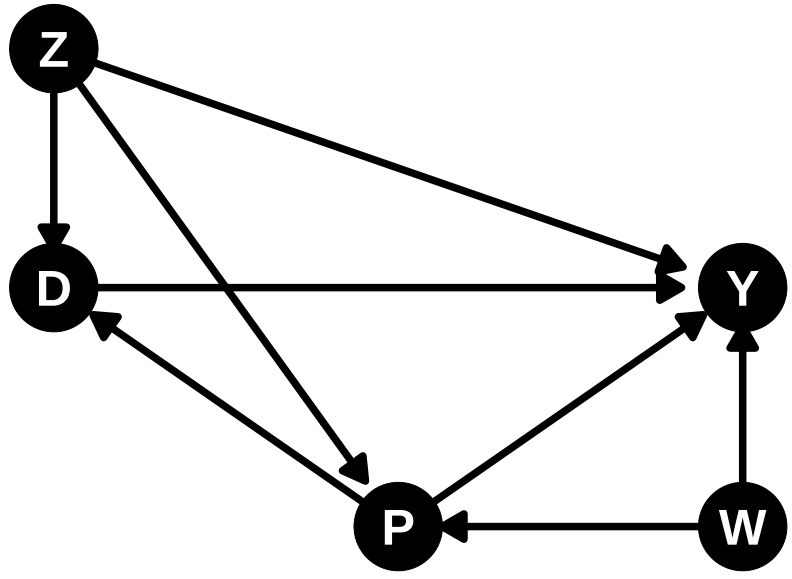
Does $\{P\}$ identify the effect of D on Y? No!

Does $\{Z, P\}$ identify the effect of D on Y? Yes!

Does $\{Z, W\}$ identify the effect of D on Y? No!

Does $\{Z, P, W\}$ identify the effect of D on Y?

SOO: Identification (DAGs)



Does $\{Z\}$ identify the effect of D on Y? No!

Does $\{P\}$ identify the effect of D on Y? No!

Does $\{Z, P\}$ identify the effect of D on Y? Yes!

Does $\{Z, W\}$ identify the effect of D on Y? No!

Does $\{Z, P, W\}$ identify the effect of D on Y? Yes!

SOO: Estimation

There are four main ways to estimate the effect of D on Y under SOO:

1. Stratification
2. Weighting
3. Matching
4. Weighting

We will have to skip 1 and 2.

SOO: Estimation (Matching)

Consider the **propensity score**:

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

It turns out that if we can make our treatment and control groups equal in terms of $\pi(X_i)$, this is sufficient for identification.

(This is actually amazing)

SOO: Estimation (Matching)

Consider the **propensity score**:

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

It turns out that if we can make our treatment and control groups equal in terms of $\pi(X_i)$, this is sufficient for identification.

(This is actually amazing)

But we have to **estimate** $\pi(X_i)$.

SOO: Estimation (Matching)

Consider the **propensity score**:

$$\pi(X_i) \equiv \Pr(D_i = 1 \mid X_i)$$

It turns out that if we can make our treatment and control groups equal in terms of $\pi(X_i)$, this is sufficient for identification.

(This is actually amazing)

But we have to **estimate** $\pi(X_i)$.

Once we do that, we can **match** on $\hat{\pi}(X_i)$ – for every treated unit, find a control unit that looks 'the same' in terms of $\hat{\pi}(X_i)$.

SOO: Estimation (Matching)

Let's estimate the propensity score using a logistic regression model.

It turns out I know that `Reactive` is the key feature that drives treatment assignment (plus noise):

```
# Fit the logistic regression (you will do this again with Ryan!)
logit_fit <- glm(Medicine_Selection ~ Reactive,
                         data = students_analysis,
                         family = binomial(link = "logit"))

# Predict the propensity scores (note "type = "responses"")
students_analysis <- students_analysis %>%
  mutate(Propensity_Score = predict(logit_fit, type = "response"))
```

SOO: Estimation (Matching)

Let's do this 'properly' with `MatchIt` (many possible bells and whistles):

```
match_result <- MatchIt::matchit(
  Medicine_Selection ~ Reactive,
  data = students_analysis,
  distance = students_analysis$Propensity_Score
  )

match_result
```

```
A `matchit` object
 - method: 1:1 nearest neighbor matching without replacement
 - distance: User-defined - number of obs.: 1000 (original), 694 (matched)
 - target estimand: ATT
 - covariates: Reactive
```

Note: If you just specify 'method = "nearest"', it will estimate the propensity score itself and use that!

SOO: Estimation (Matching)

And now estimate the ATT:

```r
matched_data <- match.data(match_result)

t.test(Headache ~ Medicine_Selection, data = matched_data)
```

```
    Welch Two Sample t-test

data:  Headache by Medicine_Selection
t = -0.73222, df = 691.98, p-value = 0.4643
alternative hypothesis: true difference in means between group 0 and group 1 is
95 percent confidence interval:
 -0.7174173  0.3276669
sample estimates:
mean in group 0 mean in group 1
       7.780188        7.975063
```

SOO: Estimation (Regression)

Finally, let's come back to our old friend, regression:

We can use regression to estimate the effect of D on Y under SOO, but we have to include X as a control:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 D_i + \hat{\beta}_2 X_i$$

Note that the interpretation of $\hat{\beta}_1$ has not changed!

Note also that $\hat{\beta}_2$ is *not* a causal effect – it is a nuisance parameter.

SOO: Estimation (Regression)

Let's do this in R:

```
lm_robust(Headache ~ Medicine_Selection + Reactive,
          data = students_analysis) %>%
  summary()
```

```
Call:
lm_robust(formula = Headache ~ Medicine_Selection + Reactive,
    data = students_analysis)

Standard error type:  HC2

Coefficients:
                  Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper  DF
(Intercept)         5.3240     0.1548 34.3996 1.352e-171   5.0203   5.6277 997
Medicine_Selection  0.1338     0.2497  0.5356  5.923e-01  -0.3563   0.6239 997
Reactive            2.5920     0.2266 11.4410  1.446e-28   2.1474   3.0366 997

Multiple R-squared:  0.1087 ,   Adjusted R-squared:  0.1069
F-statistic: 86.48 on 2 and 997 DF,  p-value: < 2.2e-16
```

SOO: Estimation (Agnostic Regression)

One well-regarded large-sample linear regression estimator (Lin, 2013):

$$Y_i = \hat{\beta}_0 + D_i\hat{\beta}_{1\text{int}} + (X_i - \bar{X})\hat{\beta}_2 + D_i(X_i - \bar{X})\hat{\beta}_3$$

Where X_i are our covariates and $\bar{X}$ is the sample mean of X_i

Read: De-mean X, and interact it with the D.

You can do this in an experiment too – you will potentially gain some efficiency advantages.

SOO: Estimation (Agnostic Regression)

The {estimatr} package gives us a canned function for this:

```
lm_lin(Headache ~ Medicine_Selection, covariates = ~ Reactive,
          data = students_analysis) %>%
  summary()
```

```
Call:
lm_lin(formula = Headache ~ Medicine_Selection, covariates = ~Reactive,
    data = students_analysis)

Standard error type:  HC2

Coefficients:
                             Estimate Std. Error  t value   Pr(>|t|) CI Lower
(Intercept)                   7.26304     0.1372 52.93796 9.100e-292   6.9938
Medicine_Selection            0.14034     0.2772  0.50631  6.128e-01  -0.4036
Reactive_c                    2.59392     0.2361 10.98517  1.407e-26   2.1306
Medicine_Selection:Reactive_c -0.03239     0.8262 -0.03921  9.687e-01  -1.6536
                             CI Upper  DF
(Intercept)                    7.5323 996
Medicine_Selection             0.6843 996
Reactive_c                     3.0573 996
Medicine_Selection:Reactive_c  1.5888 996

Multiple R-squared:  0.1087 ,   Adjusted R-squared:  0.106
F-statistic: 57.54 on 3 and 996 DF,  p-value: < 2.2e-16
```

Wrapping Up

Today we covered:

1. Fundamental building blocks for causal inference
2. How to 'solve' the FPCI with experiments
3. How to 'solve' the FPCI with conditioning

Remember: Both approaches are *assumption* driven!

Tomorrow:

1. Instrumental variables
2. Regression discontinuity
3. Difference-in-differences