Lecture 6: Causal Inference - Part 2

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

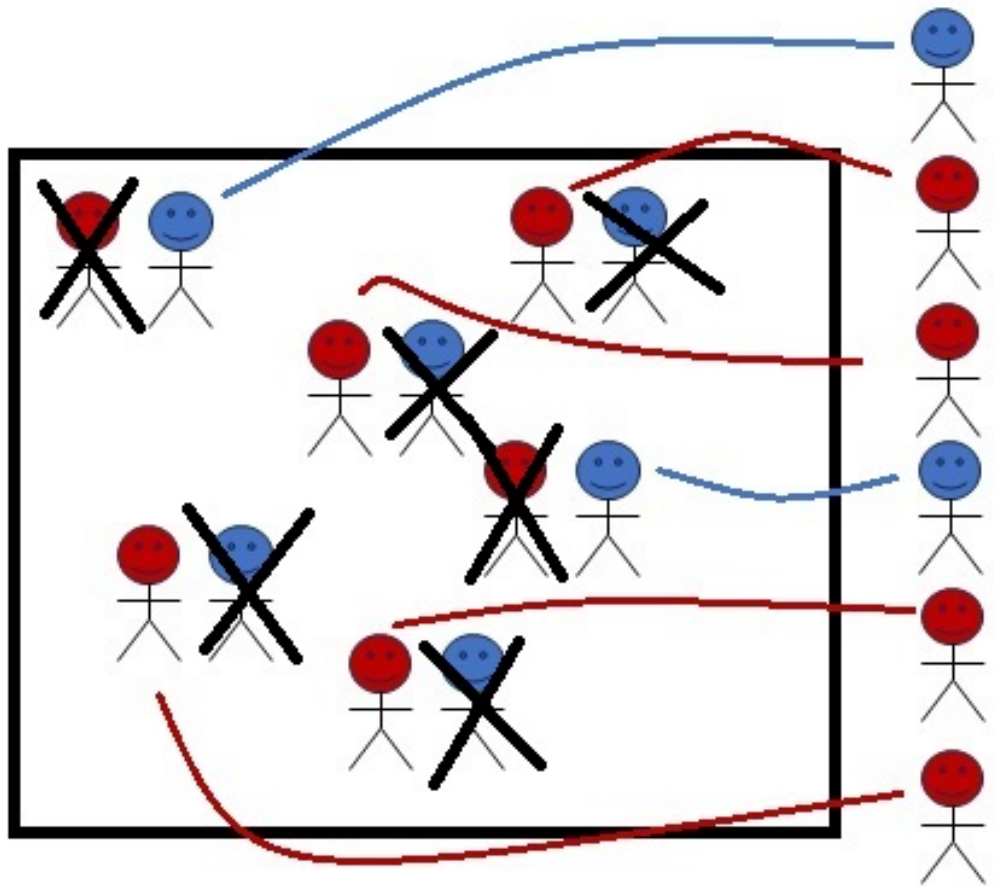
2025-07-22

Daniel de Kadt

Slides last updated: 22 July 2025

2 Causal 2 Inference

The Fundamental Problem of Causal Inference



Solutions: Assumptions + Data

Yesterday we learned about:

1. Experiments – when you get to control the assignment of D
2. SOO – when you assume the DGP of D and Y, and model them

Today:

3. Instrumental Variables
4. Regression Discontinuity
5. Difference-in-Differences

Instrumental Variables

IV: Motivation

Let's think about Squid Game again (last time, promise).

We want to estimate the effect of watching Squid Game (D) on churn (Y).

IV: Motivation

Let's think about Squid Game again (last time, promise).

We want to estimate the effect of watching Squid Game (D) on churn (Y).

Recall we had a problem: Squid Game was released to *everyone* simultaneously.

So any variation in D will have **selection** problems.

IV: Motivation

Let's think about Squid Game again (last time, promise).

We want to estimate the effect of watching Squid Game (D) on churn (Y).

Recall we had a problem: Squid Game was released to *everyone* simultaneously.

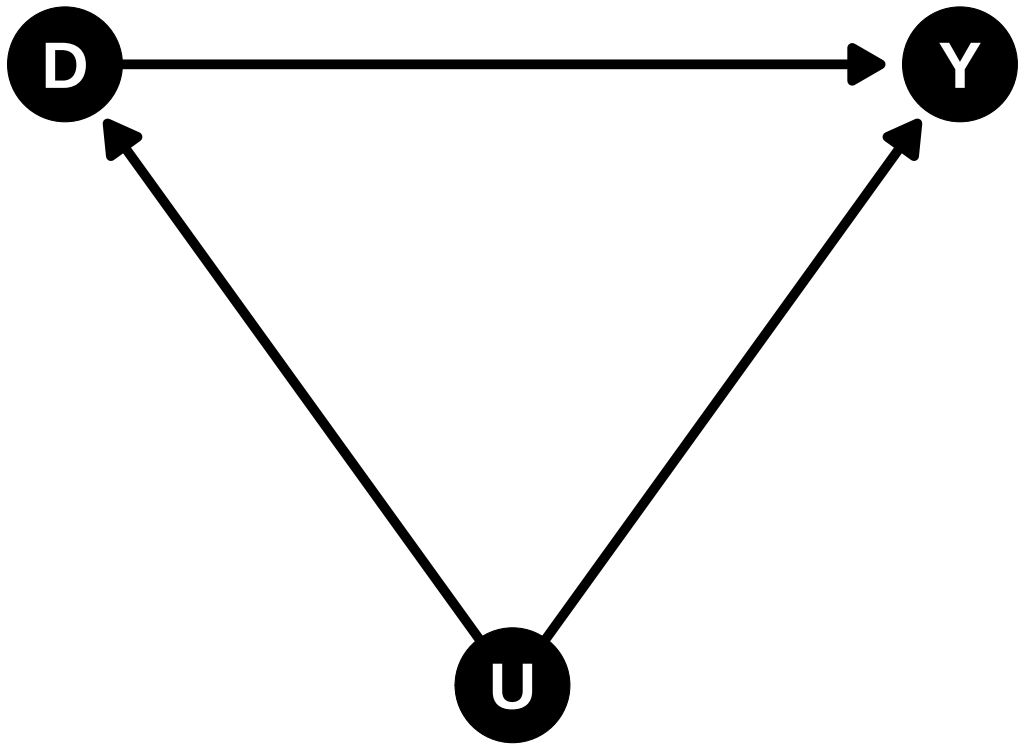
So any variation in D will have **selection** problems.

What to do?

(This example is derived from Spotify)

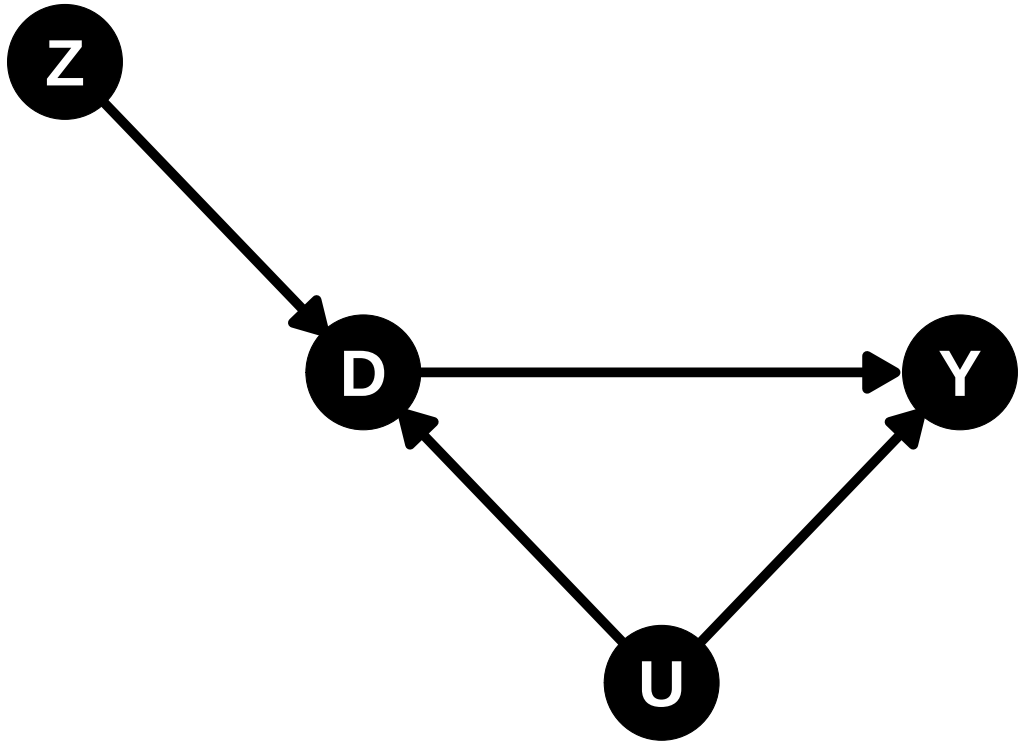
IV: Motivation

In DAG terms, our Squid Game problem looks like this: U is a canonical confounder driving selection into the show and churn.



IV: Motivation

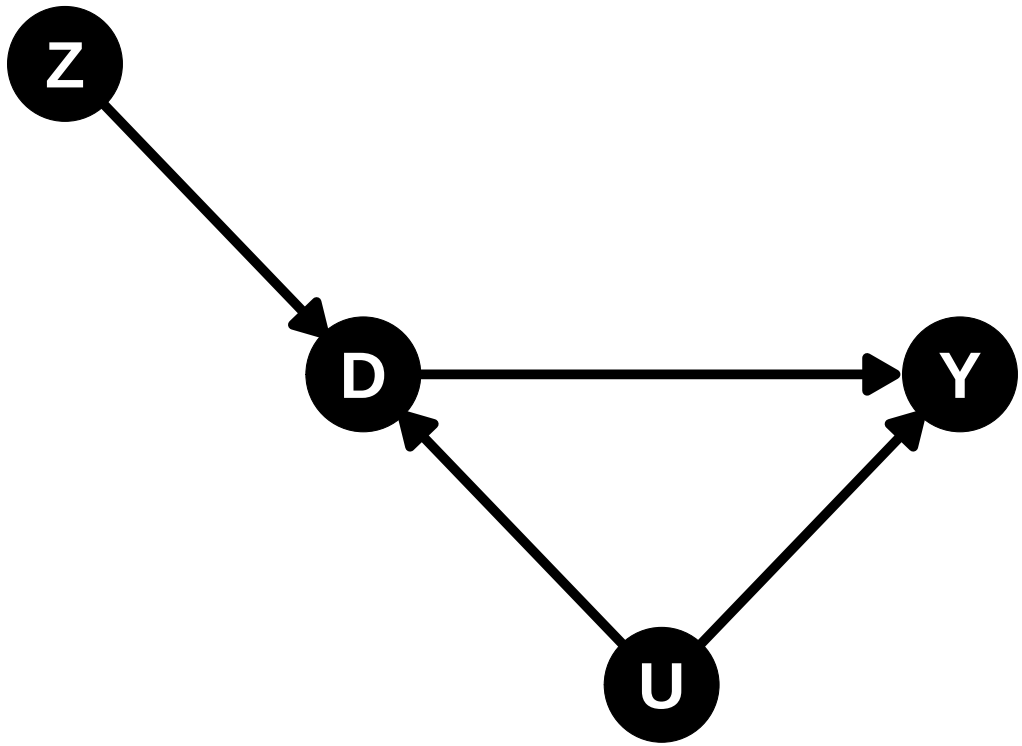
But what if we could find some Z that affects D, like so:



For example, maybe we can randomize a banner advertising Squid Game on your home page.

IV: Motivation

But what if we could find some Z that affects D, like so:



For example, maybe we can randomize a banner advertising Squid Game on your home page.

Intuition: We will use only the variation in D **induced by** Z to study the effect of D on Y.

IV: Setup

Let's start with our building blocks:

- D is a binary treatment
- Y is a continuous outcome

We want to know the effect of D on Y.

But the relationship between D and Y is confounded in some way.

We will refer to D as the 'endogenous regressor.'

IV: Setup

Consider now an **encouragement** or **instrument**: $Z_i \in \{0, 1\}$

IV: Setup

Consider now an **encouragement** or **instrument**: $Z_i \in \{0, 1\}$

Treatment potential outcomes under $Z = z$: $D_{zi} \in \{D_{1i}, D_{0i}\}$

- ➜ $D_{zi} = 1$: would receive the treatment if $Z_i = z$
- ➜ $D_{zi} = 0$: would not receive the treatment if $Z_i = z$
- ➜ e.g., $D_{1i} = 1$ encouraged to take treatment and takes treatment

Note: encouragement $\neq$ treatment

Instead: treatment $= f$(encouragement)

IV: Compliance Types

Compliance: Whether a unit follows the encouragement Z_i.

Given our setup, we can define four **compliance types**:

- Unit i is a **complier** if: $D_{1i} = 1$ and $D_{0i} = 0$
- And a **non-complier** of type:
  - **Always-takers**: $D_{1i} = D_{0i} = 1$
  - **Never-takers**: $D_{1i} = D_{0i} = 0$
  - **Defiers**: $D_{1i} = 0$ and $D_{0i} = 1$

IV: Compliance Types

Or, written as **principal strata**:

	$Z_i = 1$	$Z_i = 0$
$D_i = 1$	Complier / Always-taker	Defier / Always-taker
$D_i = 0$	Defier / Never-taker	Complier / Never-taker

IV: Potential and Realized Outcomes

Note that so far we haven't really considered Y at all.

Outcome potential outcomes: $Y_{(Z_i, D_{Z_i i})i}$

IV: Potential and Realized Outcomes

Note that so far we haven't really considered Y at all.

Outcome potential outcomes: $Y_{(Z_i, D_{Z_i i})i}$

What is observed in a given trial?

- ➜ Observed treatment indicator: $D_i = D_{Z_i i}$ for $Z_i = z$
- ➜ Observed outcome of Y_i: $Y_i = Y_{(Z_i, D_{Z_i i})i}$ for $Z_i = z$
- ➜ Thus observed outcome of Y_i can also be written as $Y_i = Y_{Z_i i}$

IV: Estimands

Intention-to-Treat (ITT):

$$ITT = \mathbb{E}[Y_{(1,D_{1i})i} - Y_{(0,D_{0i})i}]$$

Read: Effect of encouragement on outcome (regardless of treatment status). We sometimes call this the **reduced form**.

Note: If there is non-compliance, self-selection into the treatment/control groups may imply $ITT \neq ATE$

IV: Estimands

Intention-to-Treat (ITT):

$$ITT = \mathbb{E}[Y_{(1,D_{1i})i} - Y_{(0,D_{0i})i}]$$

Read: Effect of encouragement on outcome (regardless of treatment status). We sometimes call this the **reduced form**.

Note: If there is non-compliance, self-selection into the treatment/control groups may imply $ITT \neq ATE$

If you randomize Z, this is called an **encouragement design**, where $\{Y_{zd}\} \perp Z$.

In that case, our **identification result** is:

$$ITT = \mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]$$

IV: Assumptions

The ITT tells us about the effect of Z on Y

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

Four key assumptions:

1. **Relevance** of the instrument:
   $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

Four key assumptions:

1. **Relevance** of the instrument:
   $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$

2. **Ignorability** of the instrument:
   $Z \perp \{Y_{zd}, D_z\}$ (sufficient for ITT)

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

Four key assumptions:

1. **Relevance** of the instrument:
   $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$

2. **Ignorability** of the instrument:
   $Z \perp \{Y_{zd}, D_z\}$ (sufficient for ITT)

3. **Exclusion restriction**:
   $Y_{1,d} = Y_{0,d}$ for $d \in \{0, 1\}$

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

Four key assumptions:

1. **Relevance** of the instrument:
   $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$

2. **Ignorability** of the instrument:
   $Z \perp \{Y_{zd}, D_z\}$ (sufficient for ITT)

3. **Exclusion restriction**:
   $Y_{1,d} = Y_{0,d}$ for $d \in \{0, 1\}$

4. **Monotonicity**:
   $D_1 \geq D_0$ ("no defiers")

IV: Assumptions

The ITT tells us about the effect of Z on Y – what about the effect of D on Y?

Four key assumptions:

1. **Relevance** of the instrument:
   $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_0 = 1)$

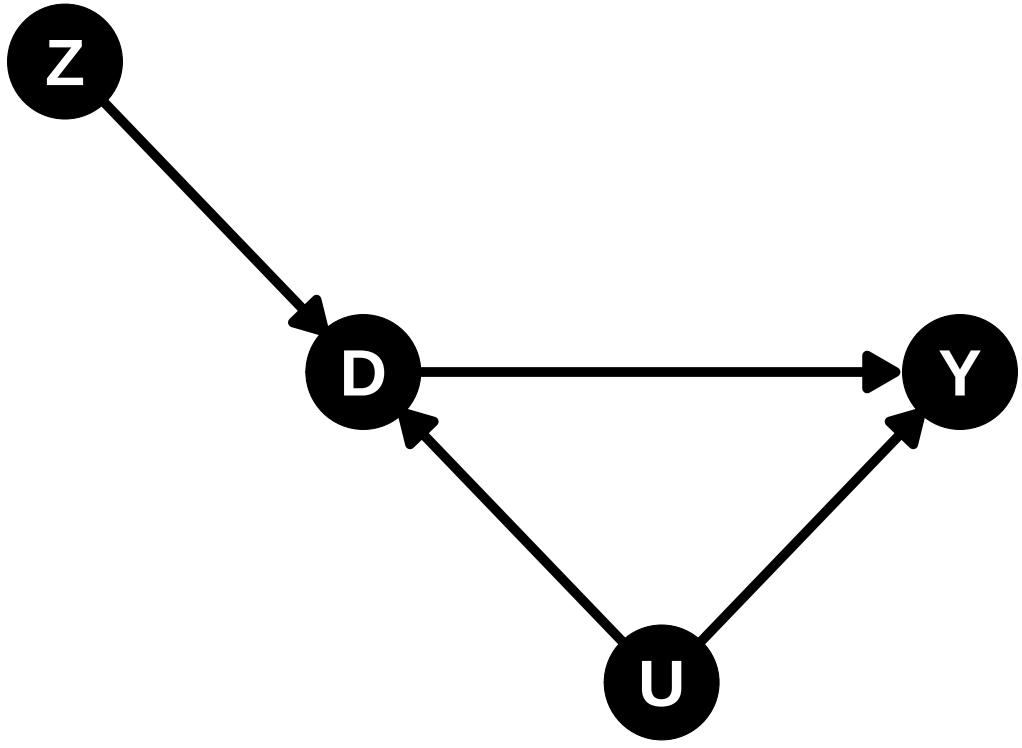
2. **Ignorability** of the instrument:
   $Z \perp \{Y_{zd}, D_z\}$ (sufficient for ITT)

3. **Exclusion restriction**:
   $Y_{1,d} = Y_{0,d}$ for $d \in \{0, 1\}$

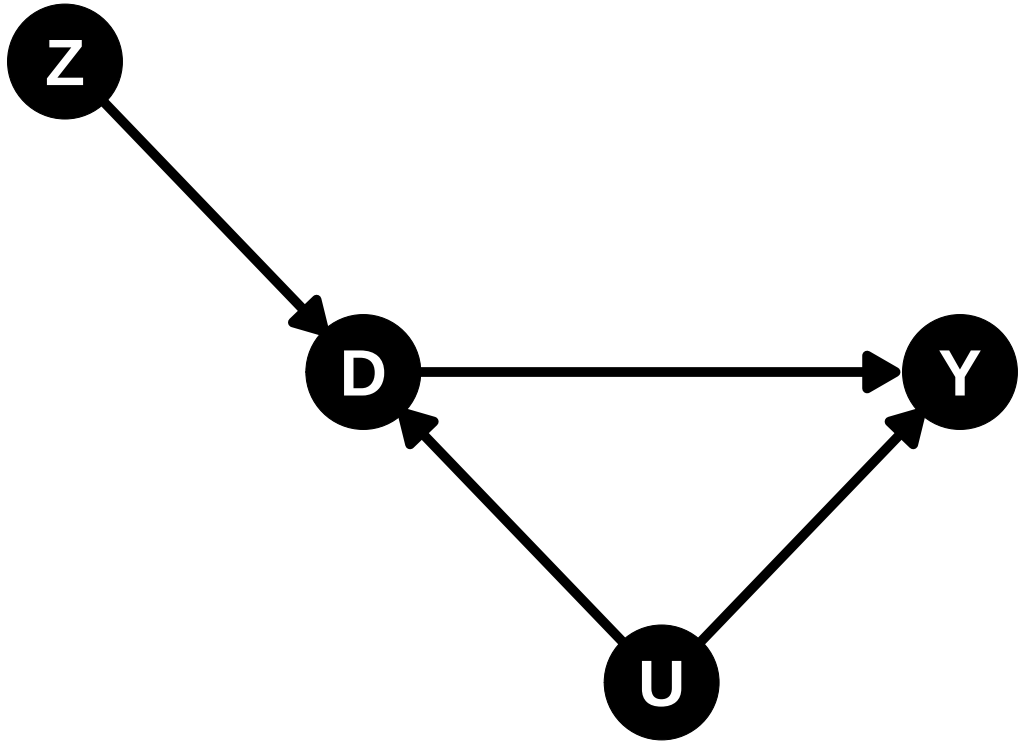
4. **Monotonicity**:
   $D_1 \geq D_0$ ("no defiers")

Intuition: Under these assumptions, we can express the **effect of D on Y** in terms of the **ITT** (which is hopefully identified).

IV: Assumptions

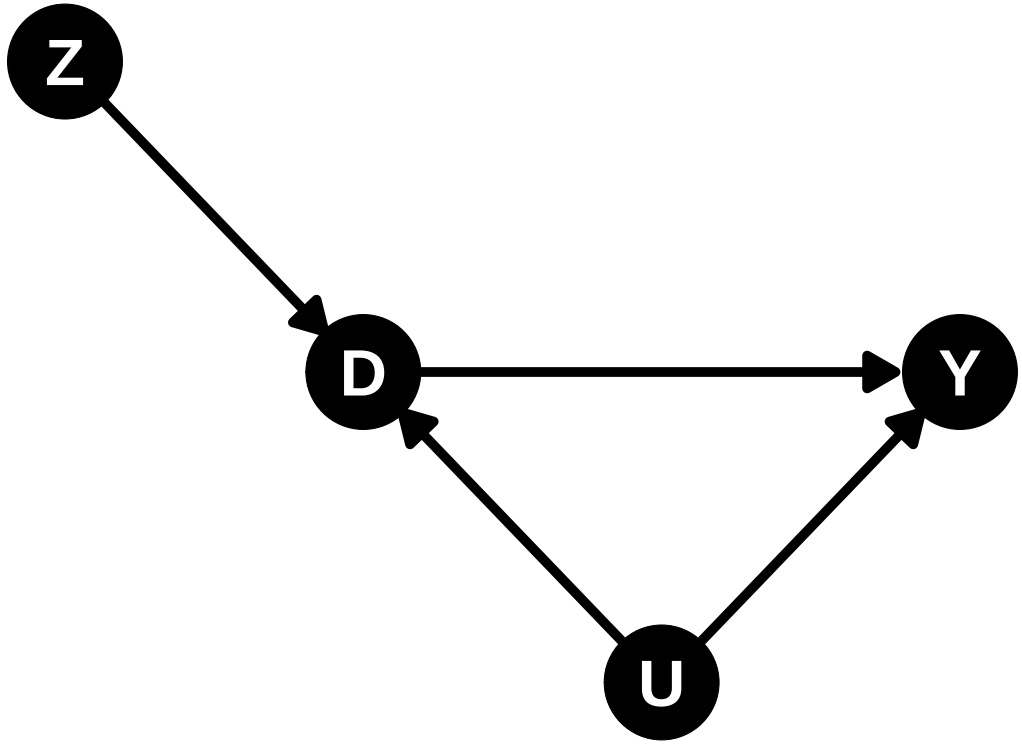


IV: Assumptions



1. Relevance: Encoded by $Z \to D$ in the DAG.

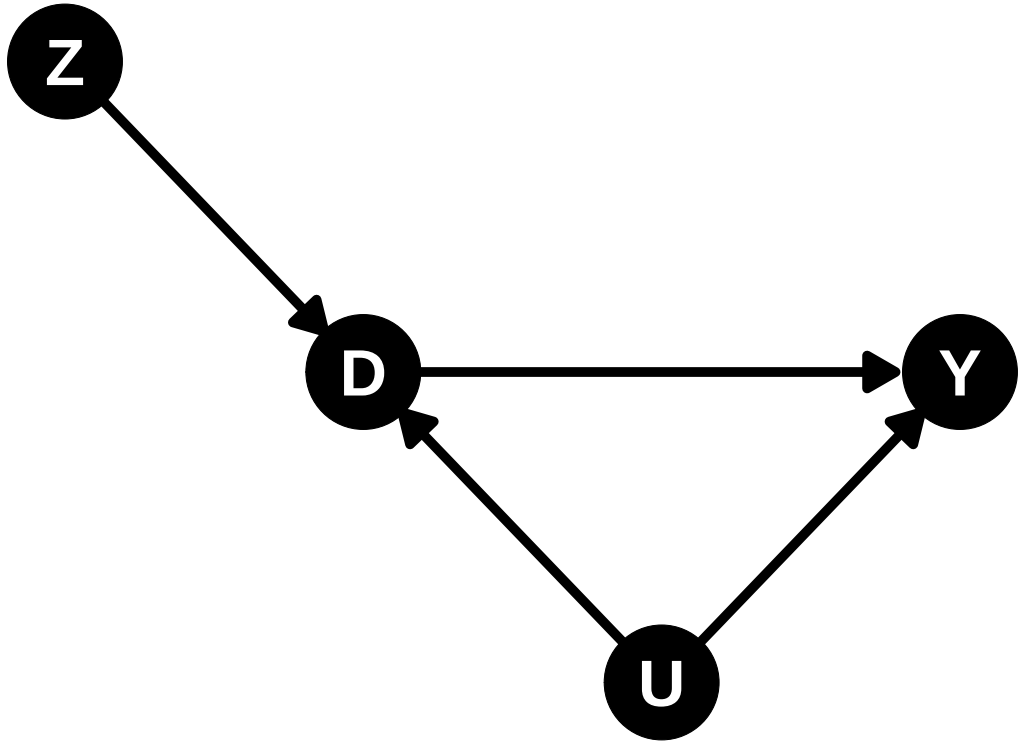
IV: Assumptions



1. Relevance: Encoded by $Z \rightarrow D$ in the DAG.

2. Ignorability: Absence of any back-door path between Z and D and between Z and Y.

IV: Assumptions



1. Relevance: Encoded by $Z \to D$ in the DAG.

2. Ignorability: Absence of any back-door path between Z and D and between Z and Y.

3. Exclusion restriction: Absence of any front-door path between Z and Y.

IV: Identification

ITT can be decomposed into a combination of subgroup-specific estimands:

$$ITT = ITT^c \times \Pr(\text{compliers}) + ITT^a \times \Pr(\text{always-takers}) + ITT^n \times \Pr(\text{never-takers}) + ITT^d \times \Pr(\text{defiers})$$

IV: Identification

ITT can be decomposed into a combination of subgroup-specific estimands:

$$ITT = ITT^c \times \Pr(\text{compliers}) + ITT^a \times \Pr(\text{always-takers}) + ITT^n \times \Pr(\text{never-takers}) + ITT^d \times \Pr(\text{defiers})$$

Where:

$$ITT^c = \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = 1, D_{0i} = 0]$$
$$ITT^a = \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = D_{0i} = 1] \quad \text{etc.}$$

IV: Identification

ITT can be decomposed into a combination of subgroup-specific estimands:

$$\begin{aligned} ITT = & ITT^c \times \Pr(\text{compliers}) + ITT^a \times \Pr(\text{always-takers}) + \\ & ITT^n \times \Pr(\text{never-takers}) + ITT^d \times \Pr(\text{defiers}) \end{aligned}$$

Where:

$$\begin{aligned} ITT^c &= \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = 1, D_{0i} = 0] \\ ITT^a &= \mathbb{E}[Y_{1i,D_{1i}} - Y_{0i,D_{0i}} \mid D_{1i} = D_{0i} = 1] \quad \text{etc.} \end{aligned}$$

Under monotonicity and exclusion, this simplifies to:

$$ITT = ITT^c \times \Pr(\text{compliers})$$

IV: Identification

Therefore, ITT^c can be **nonparametrically identified**:

$$ITT^c = \frac{ITT}{\Pr(\text{compliers})} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]}$$

IV: Identification

Therefore, ITT^c can be **nonparametrically identified**:

$$ITT^c = \frac{ITT}{\Pr(\text{compliers})} = \frac{\mathbb{E}[Y_i \mid Z_i = 1] - \mathbb{E}[Y_i \mid Z_i = 0]}{\mathbb{E}[D_i \mid Z_i = 1] - \mathbb{E}[D_i \mid Z_i = 0]}$$

ITT^c is called the **Local Average Treatment Effect (LATE)** for compliers:

$$ITT^c = LATE = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_{1i} = 1, D_{0i} = 0]$$

IV: Estimation

LATE has a clear **causal interpretation**, but it raises important questions:

- ➜ How do we generalize from compliers to the entire population?
- ➜ Are compliers even interesting?
- ➜ We can never identify individual compliers – we only know the group average.
- ➜ Different encouragements (instruments) may identify different complier groups (uh oh).

IV: Estimation

How should we estimate LATE?

Option 1: A **plug-in estimator** called the **Wald estimator**:

$$\widehat{\tau_{LATE}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} Z_i Y_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)Y_i}{\frac{1}{n_1}\sum_{i=1}^{n} Z_i D_i - \frac{1}{n_0}\sum_{i=1}^{n}(1-Z_i)D_i} = \frac{\widehat{\text{Cov}}(Y_i, Z_i)}{\widehat{\text{Cov}}(D_i, Z_i)}$$

Where: $n_1 = \sum_{i=1}^{n} D_i$, $n_0 = n - n_1$

IV: Estimation

Option 2: A **two-stage least squares (2SLS)** estimator:

We assume two **DGPs** for the potential outcomes:

1. **First stage**: $D_z = \mu + \rho Z + \eta$
2. **Second stage**: $Y_{zd} = \gamma + \alpha D + \epsilon$

2SLS estimates these via two OLS steps:

- ➜ **Stage 1**: Regress D on Z and obtain fitted values $\hat{D}_i$
- ➜ **Stage 2**: Regress Y on $\hat{D}_i$

In **R**, 2SLS can be implemented with:

- ➜ `lm()` (manually, but SEs need correction)
- ➜ `AER::ivreg` (handles SEs properly)

IV in Action

Let's consider some simulated Squid Game viewership data:

	ID	Boredom	Income	Banner_Ad_SG	Watched_SG	Churned_6mo
1	1	-0.31423130	47187.00	0	1	1
2	2	1.87160961	40549.78	0	0	1
3	3	1.26168761	51694.72	0	1	0
4	4	0.85594984	51766.90	0	0	1
5	5	0.72861292	77088.54	0	0	1
6	6	-0.04080788	28675.99	0	0	0

IV in Action

First, we can estimate the **reduced form** (ITT) of the effect of the banner ad on churn:

```
reduced_form <- mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 1]) -
                mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 0])
reduced_form
```

```
[1] -0.08731433
```

IV in Action

First, we can estimate the **reduced form** (ITT) of the effect of the banner ad on churn:

```
reduced_form <- mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 1]) -
                mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 0])
reduced_form
```

```
[1] -0.08731433
```

Now, we need the **first stage** or pr(compliers):

```
first_stage <- mean(sg_data$Watched_SG[sg_data$Banner_Ad_SG == 1]) -
               mean(sg_data$Watched_SG[sg_data$Banner_Ad_SG == 0])
first_stage
```

```
[1] 0.1928944
```

IV in Action

First, we can estimate the **reduced form** (ITT) of the effect of the banner ad on churn:

```
reduced_form <- mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 1]) -
                mean(sg_data$Churned_6mo[sg_data$Banner_Ad_SG == 0])
reduced_form
```

```
[1] -0.08731433
```

Now, we need the **first stage** or pr(compliers):

```
first_stage <- mean(sg_data$Watched_SG[sg_data$Banner_Ad_SG == 1]) -
               mean(sg_data$Watched_SG[sg_data$Banner_Ad_SG == 0])
first_stage
```

```
[1] 0.1928944
```

Finally, we can estimate the **LATE** using the Wald estimator:

```
reduced_form/first_stage
```

```
[1] -0.4526535
```

IV in Action

We can also estimate the LATE using **2SLS**:

```
AER::ivreg(Churned_6mo ~ Watched_SG | Banner_Ad_SG, data = sg_data) %>%
  summary()
```

```
Call:
AER::ivreg(formula = Churned_6mo ~ Watched_SG | Banner_Ad_SG,
    data = sg_data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4629 -0.4629 -0.0102  0.5371  0.9898

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.46285    0.06238   7.420  2.5e-13 ***
Watched_SG  -0.45265    0.14642  -3.091  0.00205 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4458 on 998 degrees of freedom
Multiple R-Squared: 0.005351,   Adjusted R-squared: 0.004355
Wald test: 9.557 on 1 and 998 DF,  p-value: 0.002047
```

IV: Good Practice

Things to pay attention to:

1. Strength of first stage (test this)
2. Assignment of and properties of Z (e.g. balance)
3. Plausibility of exclusion restriction (theory)
4. Focus on interpration (e.g. compliers)
5. Characterize the compliers if you can!

Regression Discontinuity

RDD: Motivation

I've just hired you at my e-commerce store. We're trying to understand the value of our promised '1-day delivery' product.

Question: Does 1-day delivery (D) affect whether a customer completes a transaction (Y).

RDD: Motivation

I've just hired you at my e-commerce store. We're trying to understand the value of our promised '1-day delivery' product.

Question: Does 1-day delivery (D) affect whether a customer completes a transaction (Y).

But promised 1-day delivery is a function of the time of day you checkout.

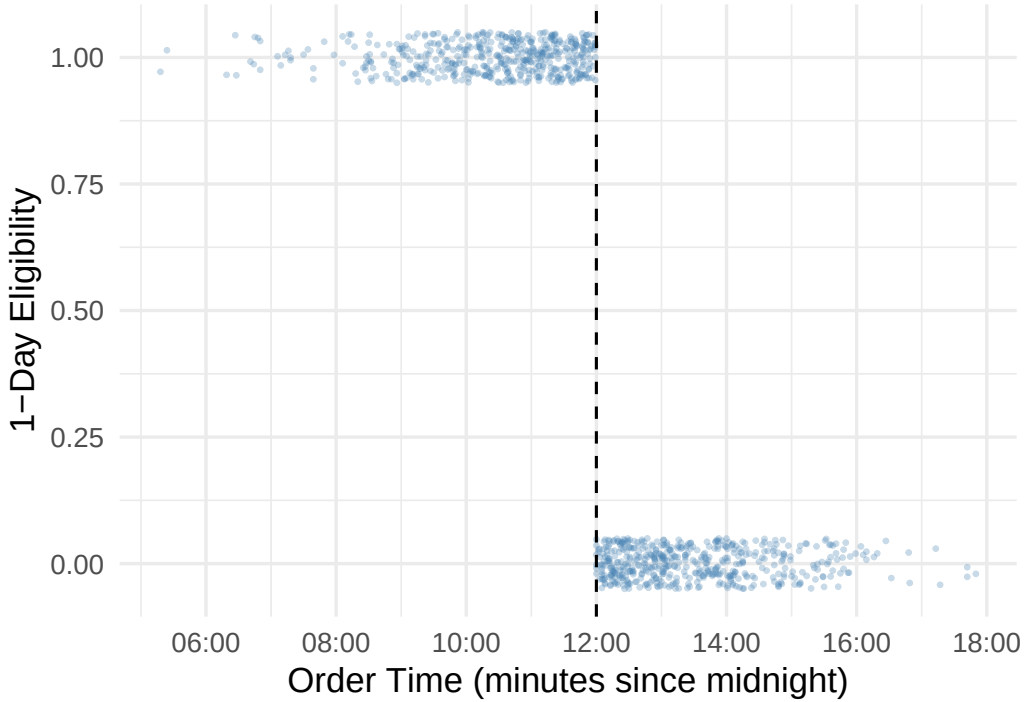
So we have a **selection** problem: when you order is not random, so neither is D!

What to do?

(This example is derived from Instacart)

RDD: Motivation

Our company has a **sharp** rule: If you checkout at 11:59am or earlier, you get 1-day. If it's 12:00pm or later, you get 2+ day delivery.



Intuition: If people who checkout at 11:59am are *exactly the same* as those who checkout at 12:00pm, then the variation in D among just those customers is independent of potential outcomes.

RDD: Setup

Formalising the RDD:

- $D_i \in \{0, 1\}$: Treatment
- X_i: **Forcing variable** (aka **running variable** or **score**) that perfectly determines D_i at cutpoint c:

$$D_i = \mathbf{1}\{X_i \geq c\} \quad \text{or equivalently} \quad D_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

- Potential outcomes: $\mathbb{E}[Y_{0i} \mid X_i]$ and $\mathbb{E}[Y_{1i} \mid X_i]$, **defined** for every value of X_i

Note: X_i may be correlated with Y_{0i} and Y_{1i}! (It likely is – why?)

RDD: Setup

Hold up – this looks kind of like **selection on observables**...

RDD: Setup

Hold up – this looks kind of like **selection on observables**...

If potential outcomes are a deterministic function of X_i, why not just **adjust or control** for X_i?

RDD: Setup

Hold up – this looks kind of like **selection on observables**...

If potential outcomes are a deterministic function of X_i, why not just **adjust or control** for X_i?

Lack of common support $\rightarrow$ across all i, only **one** of Y_{0i} and Y_{1i} can be observed for each level of X_i.

RDD: Identification

Intuition: suppose there is no discontinuity in **potential outcomes** $\mathbb{E}[Y_{0i} \mid X_i = x]$ and $\mathbb{E}[Y_{1i} \mid X_i = x]$ at the threshold c.

If $\mathbb{E}[Y_{0i} \mid X_i = x]$ and $\mathbb{E}[Y_{1i} \mid X_i = x]$ can be approximated by some $f(X_i)$, estimate missing potential outcomes by **extrapolating** to $X_i = c$.

RDD: Identification

Intuition: suppose there is no discontinuity in **potential outcomes** $\mathbb{E}[Y_{0i} \mid X_i = x]$ and $\mathbb{E}[Y_{1i} \mid X_i = x]$ at the threshold c.

If $\mathbb{E}[Y_{0i} \mid X_i = x]$ and $\mathbb{E}[Y_{1i} \mid X_i = x]$ can be approximated by some $f(X_i)$, estimate missing potential outcomes by **extrapolating** to $X_i = c$.

Any difference in Y_i at $X_i = c$ is a **causal effect**!

Estimand: Local Average Treatment Effect (LATE) at the threshold

$$LATE_{SRD} = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

RDD: Assumption

Continuity of average potential outcomes:

$$\lim_{\epsilon \uparrow 0} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{0i} \mid X_i = c]$$

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{1i} \mid X_i = c]$$

Read: Potential outcomes arbitrarily close to c are approximately the same as potential outcomes exactly at c.

RDD: Assumption

Continuity of average potential outcomes:

$$\lim_{\epsilon \uparrow 0} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{0i} \mid X_i = c]$$

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{1i} \mid X_i = c]$$

Read: Potential outcomes arbitrarily close to c are approximately the same as potential outcomes exactly at c.

A simple proof:

$$\lim_{\epsilon \downarrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon]$$

RDD: Assumption

Continuity of average potential outcomes:

$$\lim_{\epsilon \uparrow 0} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{0i} \mid X_i = c]$$

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{1i} \mid X_i = c]$$

Read: Potential outcomes arbitrarily close to c are approximately the same as potential outcomes exactly at c.

A simple proof:

$$\begin{aligned} &\lim_{\epsilon \downarrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] \\ =&\lim_{\epsilon \downarrow c} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] \end{aligned}$$

RDD: Assumption

Continuity of average potential outcomes:

$$\lim_{\epsilon \uparrow 0} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{0i} \mid X_i = c]$$

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{1i} \mid X_i = c]$$

Read: Potential outcomes arbitrarily close to c are approximately the same as potential outcomes exactly at c.

A simple proof:

$$\begin{aligned}
&\lim_{\epsilon \downarrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] \\
=&\lim_{\epsilon \downarrow c} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] \\
=&\mathbb{E}[Y_{1i} \mid X_i = c] - \mathbb{E}[Y_{0i} \mid X_i = c] \quad \text{(by continuity)}
\end{aligned}$$

RDD: Assumption

Continuity of average potential outcomes:

$$\lim_{\epsilon \uparrow 0} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{0i} \mid X_i = c]$$

$$\lim_{\epsilon \downarrow 0} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] = \mathbb{E}[Y_{1i} \mid X_i = c]$$

Read: Potential outcomes arbitrarily close to c are approximately the same as potential outcomes exactly at c.

A simple proof:

$$\begin{aligned}
&\lim_{\epsilon \downarrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_i \mid X_i = c + \epsilon] \\
=&\lim_{\epsilon \downarrow c} \mathbb{E}[Y_{1i} \mid X_i = c + \epsilon] - \lim_{\epsilon \uparrow c} \mathbb{E}[Y_{0i} \mid X_i = c + \epsilon] \\
=&\mathbb{E}[Y_{1i} \mid X_i = c] - \mathbb{E}[Y_{0i} \mid X_i = c] \quad \text{(by continuity)} \\
=&\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c] = LATE_{SRD}
\end{aligned}$$

QED

RDD: Estimation (Linear Regression)

In the continuity framework, estimation is an **extrapolation problem**.

One very simple approach would be to assume a parametric DGP, where *LATE* is constant, and potential outcomes are linear in X_i:

$$Y_{di} = \alpha + LATE_{SRD} \cdot d + \beta X_i$$

Note: This is an *assumption* about the DGP – it may well be wrong.

RDD: Estimation (Linear Regression)

Given the assumed DGP we just saw, a very reasonable estimator for the *LATE* would be linear regression:

RDD: Estimation (Linear Regression)

Given the assumed DGP we just saw, a very reasonable estimator for the *LATE* would be linear regression:

1. Recenter forcing variable: $\tilde{X}_i = X_i - c$

RDD: Estimation (Linear Regression)

Given the assumed DGP we just saw, a very reasonable estimator for the *LATE* would be linear regression:

1. Recenter forcing variable: $\tilde{X}_i = X_i - c$
2. Regress: $Y_i = \hat{\alpha} + \hat{LATE} D_i + \hat{\beta}\tilde{X}_i$

RDD: Estimation (Linear Regression)

Given the assumed DGP we just saw, a very reasonable estimator for the *LATE* would be linear regression:

1. Recenter forcing variable: $\tilde{X}_i = X_i - c$
2. Regress: $Y_i = \hat{\alpha} + \hat{LATE} D_i + \hat{\beta}\tilde{X}_i$
3. $\hat{LATE}$ is an unbiased estimator of the LATE

We could assume a more **flexible** (realistic?) functional form, e.g., varying slopes in X_i, or polynomial functions of X_i, and fit that regression too.

RDD: Estimation (Local Polynomial Approximation)

Whatever function we choose, we make strong parametric assumptions.

Current state of the art is **local polynomial approximation**, which offers a **non-parametric** estimator of $LATE_{SRD}$.

Proceeds as follows:

1. Choose **bandwidth** or window h
2. Choose **polynomial** order p and **kernel** function $K(\cdot)$
3. Fit two weighted regressions (for $X_i \geq c$ and $X_i < c$) on either side of c to estimate two intercepts: $\mu_{\downarrow}$ and $\mu_{\uparrow}$
4. Calculate $\hat{LATE}_{SRD} = \hat{\mu}_{\downarrow} - \hat{\mu}_{\uparrow}$

Implemented with `rdrobust()` in **R**.

RDD in Action

First, let's clean our data and estimate the effect of 1-day delivery on completion using linear regression:

```
# Some data cleaning we need to do:
delivery_data <- delivery_data %>%
  mutate(Order_Minutes_Rev = 1439 - Order_Minutes, # Reverse the minutes
         OMR_c = Order_Minutes_Rev - 720) # Recenter forcing variable

# Now we can estimate the LATE using linear regression:
lm_robust(Order_Placed ~ OneDay + OMR_c, data = delivery_data) %>%
  summary()
```

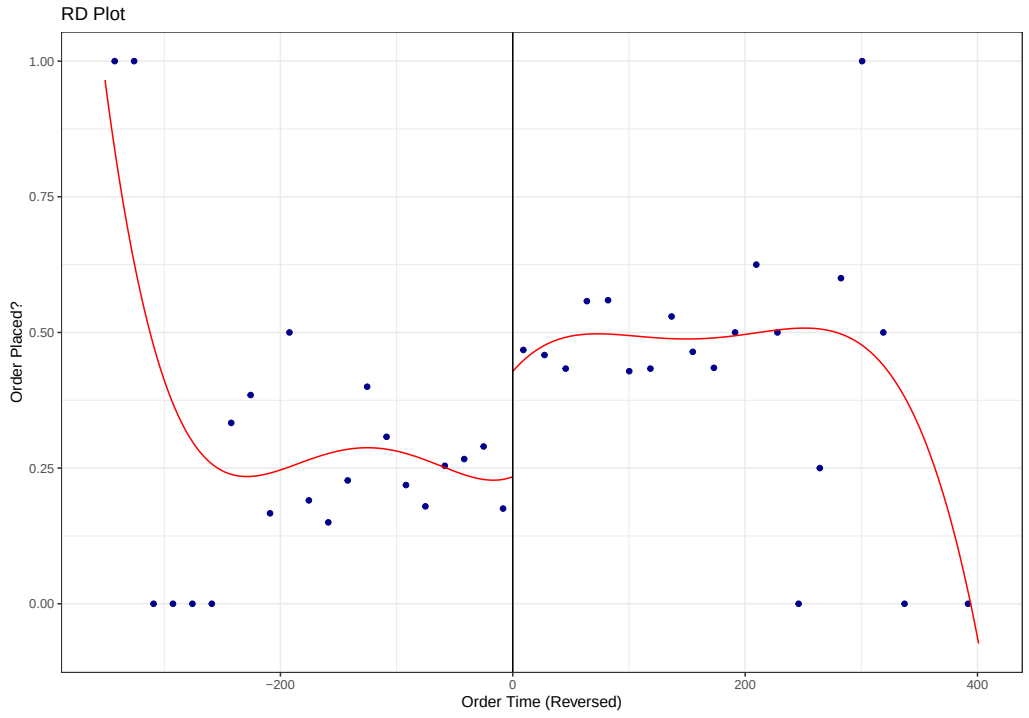
```
Call:
lm_robust(formula = Order_Placed ~ OneDay + OMR_c, data = delivery_data)

Standard error type:  HC2

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)   CI Lower  CI Upper  DF
(Intercept)  0.2403031   0.027032   8.890 2.821e-18  0.1872569 0.2933493 997
OneDay       0.2636326   0.049136   5.365 1.005e-07  0.1672103 0.3600550 997
OMR_c       -0.0002181   0.000211  -1.034 3.015e-01 -0.0006322 0.0001959 997

Multiple R-squared:  0.05407 ,  Adjusted R-squared:  0.05217
F-statistic: 28.64 on 2 and 997 DF,  p-value: 8.073e-13
```

RDD in Action

Now let's visualize the RDD effect using `rdplot` from {rdrobust}:

```
rdplot(y = delivery_data$Order_Placed, x = delivery_data$OMR_c,
       c = 0,
       x.label = "Order Time (Reversed)",
       y.label = "Order Placed?") # Bandwidth
```

```
[1] "Mass points detected in the running variable."
```



RDD in Action

Use `rdrobust()` to implement local polynomial approx.:

```
suppressWarnings(
  rdrobust(y = delivery_data$Order_Placed,
        x = delivery_data$OMR_c,
        c = 0)) %>%
  summary()
```

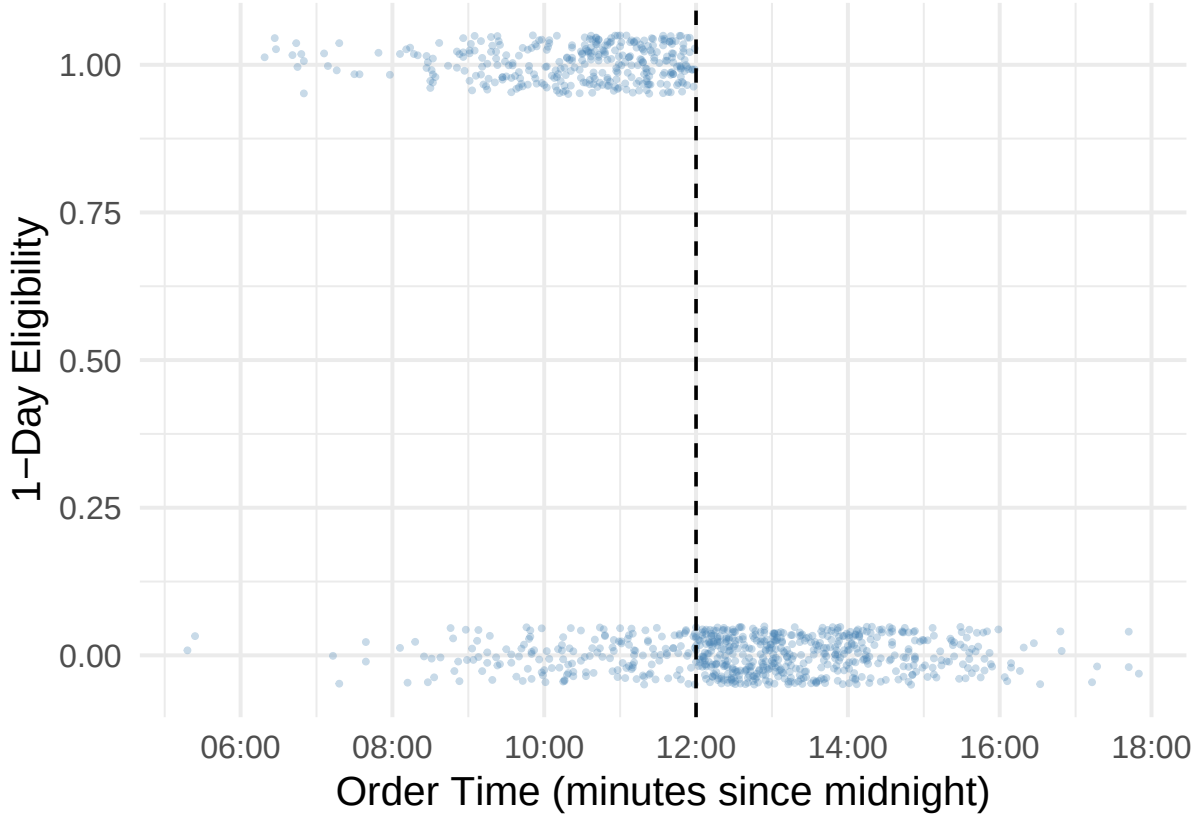
```
Sharp RD estimates using local polynomial regression.

Number of Obs.                 1000
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 503         497
Eff. Number of Obs.            281         275
Order est. (p)                   1           1
Order bias  (q)                  2           2
BW est. (h)                 89.274      89.274
BW bias (b)                139.518     139.518
rho (h/b)                    0.640       0.640
Unique Obs.                    204         206

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     0.239     0.083     2.878     0.004     [0.076 , 0.402]
        Robust         -         -     2.402     0.016     [0.044 , 0.431]
=============================================================================
```

RDD: Good Practice

1. Check that your setting is actually an RDD (!)
2. Visualise the jump in Y as a function of X – be very skeptical of these plots!
3. Check for discontinuities in background covariates (e.g. lagged Y is great)
4. Check 'placebo' discontinuities (other parts of X) – but do this on *either side* of c separately
5. Generally stick with `rdrobust()` defaults. But also see `rdhonest`.

RDD Meets IV: Getting Fuzzy

What if, instead of our company having a sharp cut-off at noon, we had a **fuzzy** cut-off?



RDD Meets IV: Getting Fuzzy

What if, instead of our company having a sharp cut-off at noon, we had a **fuzzy** cut-off?

What's going on? If you order late, you will *never* get 1-day delivery, but if you order early, you might (or might not) get it.

RDD Meets IV: Getting Fuzzy

This looks like IV!

Formalising this research setting:

- $Z_i \in \{0, 1\}$: Encouragement

RDD Meets IV: Getting Fuzzy

This looks like IV!

Formalising this research setting:

- $Z_i \in \{0, 1\}$: Encouragement
- $D_i \in \{0, 1\}$: Treatment, a probabilistic function of Z_i

RDD Meets IV: Getting Fuzzy

This looks like IV!

Formalising this research setting:

- $Z_i \in \{0, 1\}$: Encouragement
- $D_i \in \{0, 1\}$: Treatment, a probabilistic function of Z_i
- X_i: Forcing variable perfectly determines Z_i with cutpoint c:

$$Z_i = \mathbf{1}\{X_i \geq c\} \quad \text{or equivalently} \quad Z_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases}$$

Note: The reduced form (effect of Z_i on Y_i) is just a **sharp RDD**!

RDD Meets IV: Getting Fuzzy

Local ITT (LITT) of encouragement at the threshold:

$$LITT = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

RDD Meets IV: Getting Fuzzy

Local ITT (LITT) of encouragement at the threshold:

$$LITT = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

LATE for compliers at the threshold

$$LATE^c = \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{unit } i \text{ is a complier and } X_i = c]$$

RDD Meets IV: Getting Fuzzy

Local ITT (LITT) of encouragement at the threshold:

$$LITT = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

LATE for compliers at the threshold

$$LATE^c = \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{unit } i \text{ is a complier and } X_i = c]$$

Assumptions:

1. 'Augmented' continuity: Both $\mathbb{E}[D_{zi} \mid X_i = x]$ (p.o. for treatment) and $\mathbb{E}[Y_{zi} \mid X_i = x]$ (p.o. for outcome) are continuous in x around $X_i = c$, for $z = 0, 1$

RDD Meets IV: Getting Fuzzy

Local ITT (LITT) of encouragement at the threshold:

$$LITT = \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

LATE for compliers at the threshold

$$LATE^c = \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{unit } i \text{ is a complier and } X_i = c]$$

Assumptions:

1. 'Augmented' continuity: Both $\mathbb{E}[D_{zi} \mid X_i = x]$ (p.o. for treatment) and $\mathbb{E}[Y_{zi} \mid X_i = x]$ (p.o. for outcome) are continuous in x around $X_i = c$, for $z = 0, 1$

2. From IV: Monotonicity, exclusion restriction, relevance of Z_i

Fuzzy RDD: Estimation

Parametric estimation for $LATE^c$:

1. Code instrument: $Z = \mathbf{1}\{X > c\}$

Fuzzy RDD: Estimation

Parametric estimation for $LATE^c$:

1. Code instrument: $Z = \mathbf{1}\{X > c\}$
2. Fit 2SLS:

$$\text{First Stage:} \quad D_i = f(X_i) + \beta Z_i + \epsilon_i$$

$$\text{Second Stage:} \quad Y_i = f(X_i) + \alpha \hat{D}_i + \nu_i$$

Note: Specification of $f(\cdot)$ is flexible but must be the same in both stages.

Fuzzy RDD: Estimation

Non-parametric estimation:

1. $LITT$ can be estimated using local polynomial approximation, as the LATE was for a sharp RDD. (Why?)

2. Proportion of compliers can likewise be estimated with D_i as the outcome

3. $LATE_c$ (for compliers at the threshold) is just:

$$LATE_c = \frac{LITT}{Pr(\text{Compliers} \mid X_i = c)}$$

Fuzzy RDD: Estimation

Non-parametric estimation:

1. *LITT* can be estimated using local polynomial approximation, as the LATE was for a sharp RDD. (Why?)

2. Proportion of compliers can likewise be estimated with D_i as the outcome

3. $LATE_c$ (for compliers at the threshold) is just:

$$LATE_c = \frac{LITT}{Pr(\text{Compliers} \mid X_i = c)}$$

Whatever you do, it is **critical** that you test and visualise the first stage. A weak (or non-existent) first stage generates severe bias, and misleads.

Fuzzy RDD in Action

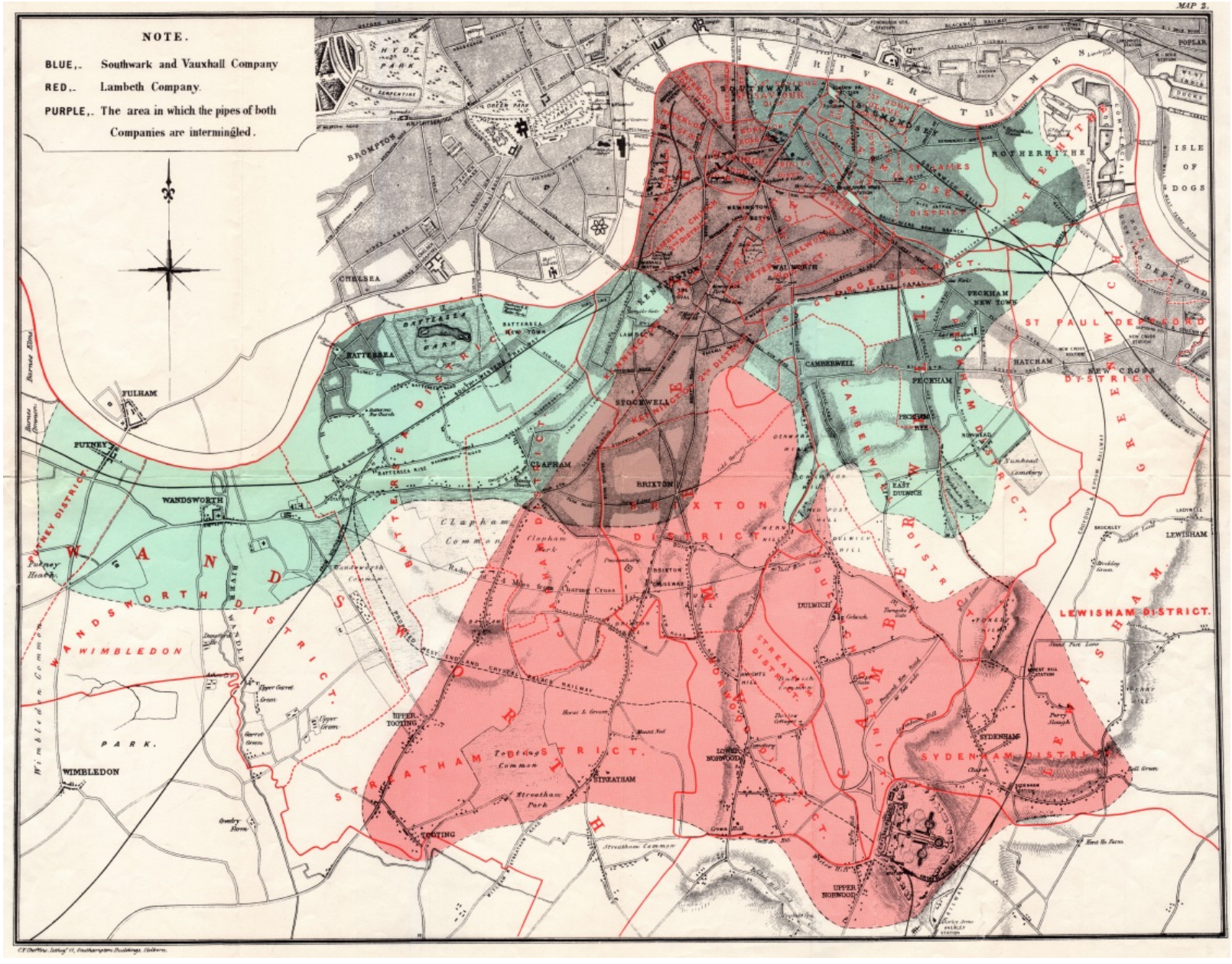
```
Fuzzy RD estimates using local polynomial regression.

Number of Obs.                 1000
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 503          497
Eff. Number of Obs.            281          275
Order est. (p)                   1            1
Order bias  (q)                  2            2
BW est. (h)                 89.274       89.274
BW bias (b)                139.518      139.518
rho (h/b)                    0.640        0.640
Unique Obs.                    204          206

First-stage estimates.

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     0.658     0.065    10.195     0.000     [0.532 , 0.785]
        Robust         -         -     8.205     0.000     [0.489 , 0.795]
=============================================================================

Treatment effect estimates.

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     0.363     0.123     2.956     0.003     [0.122 , 0.604]
        Robust         -         -     2.525     0.012     [0.083 , 0.657]
=============================================================================
```

Difference-in-Differences

DiD: Motivation

DiD: Motivation

DiD: Motivation



DiD: Motivation

TABLE XII.

Sub-Districts.	Deaths from Cholera in 1849.	Deaths from Cholera in 1854.	Water Supply.
St. Saviour, Southwark	283	371	
St. Olave	157	161	
St. John, Horsleydown	192	148	
St. James, Bermondsey	249	362	
St. Mary Magdalen	259	244	
Leather Market	226	237	
Rotherhithe*	352	282	Southwark & Vauxhall Company only.
Wandsworth	97	59	
Battersea	111	171	
Putney	8	9	
Camberwell	235	240	
Peckham	92	174	
Christchurch, Southwark	256	113	
Kent Road	267	174	
Borough Road	312	270	
London Road	257	93	
Trinity, Newington	318	210	
St. Peter, Walworth	446	388	Lambeth Company, and Southwark and Vauxhall Compy.
St. Mary, Newington	143	92	
Waterloo Road (1st)	193	58	
Waterloo Road (2nd)	243	117	
Lambeth Church (1st)	215	49	
Lambeth Church (2nd)	544	193	
Kennington (1st)	187	303	
Kennington (2nd)	153	142	
Brixton	81	48	
Clapham	114	165	
St. George, Camberwell	176	132	
Norwood	2	10	
Streatham	154	15	Lambeth Company only.
Dulwich	1	—	
Sydenham	5	12	
First 12 sub-districts	2261	2458	Southwk. & Vauxhall.
Next 16 sub-districts	3905	2547	Both Companies.
Last 4 sub-districts	162	37	Lambeth Company.

DiD: Motivation

Table 1 John Snow's data on mortality from cholera in areas served by only one of the Southwark & Vauxhall or Lambeth Water Companies before and after the change in water source for the Lambeth Water Company. Rates for all time points were calculated based on 1851 population census

Water supply	Cholera deaths, 1849, rate per 100,000	Cholera deaths, 1854, rate per 100,000
Lambeth Company Only	847	193

DiD: Motivation

Table 1 John Snow's data on mortality from cholera in areas served by only one of the Southwark & Vauxhall or Lambeth Water Companies before and after the change in water source for the Lambeth Water Company. Rates for all time points were calculated based on 1851 population census

Water supply	Cholera deaths, 1849, rate per 100,000	Cholera deaths, 1854, rate per 100,000	Difference in rates comparing 1854 to 1849, rate per 100,000
Lambeth Company Only	847	193	−653

DiD: Motivation

Table 1 John Snow's data on mortality from cholera in areas served by only one of the Southwark & Vauxhall or Lambeth Water Companies before and after the change in water source for the Lambeth Water Company. Rates for all time points were calculated based on 1851 population census

Water supply	Cholera deaths, 1849, rate per 100,000	Cholera deaths, 1854, rate per 100,000	Difference in rates comparing 1854 to 1849, rate per 100,000
Southwark & Vauxhall Company only	1349	1466	118
Lambeth Company Only	847	193	−653

DiD: Motivation

Table 1 John Snow's data on mortality from cholera in areas served by only one of the Southwark & Vauxhall or Lambeth Water Companies before and after the change in water source for the Lambeth Water Company. Rates for all time points were calculated based on 1851 population census

Water supply	Cholera deaths, 1849, rate per 100,000	Cholera deaths, 1854, rate per 100,000	Difference in rates comparing 1854 to 1849, rate per 100,000
Southwark & Vauxhall Company only	1349	1466	118
Lambeth Company Only	847	193	−653
Difference-in-difference, Lambeth versus Southwark & Vauxhall			−771

DiD: Motivation

Table 1 John Snow's data on mortality from cholera in areas served by only one of the Southwark & Vauxhall or Lambeth Water Companies before and after the change in water source for the Lambeth Water Company. Rates for all time points were calculated based on 1851 population census

Water supply	Cholera deaths, 1849, rate per 100,000	Cholera deaths, 1854, rate per 100,000	Difference in rates comparing 1854 to 1849, rate per 100,000
Southwark & Vauxhall Company only	1349	1466	118
Lambeth Company Only	847	193	−653
Difference-in-difference, Lambeth versus Southwark & Vauxhall	502	1273	−771

DiD: Setup

Units: $i \in \{1, ..., N\}$

Time periods: $t \in \{0 \text{ (pre-treatment)}, 1 \text{ (post-treatment)}\}$

Group indicator:

$$G_i = \begin{cases} 1 & \text{(treatment group)} \\ 0 & \text{(control group)} \end{cases}$$

DiD: Setup

Units in the treatment group receive treatment in $t = 1$, so:

Treatment indicator: $Z_{it} \in \{0, 1\}$

Group	$t = 0$	$t = 1$
$G_i = 1$ (treatment)	$Z_{i0} = 0$ (untreated)	$Z_{i1} = 1$ (treated)
$G_i = 0$ (control)	$Z_{i0} = 0$ (untreated)	$Z_{i1} = 0$ (untreated)

DiD: Setup

Define **potential outcomes** $Y_{it}(z)$ as:

- $Y_{it}(0)$: potential outcome for i in period t when untreated
- $Y_{it}(1)$: potential outcome for i in period t when treated

Note: Pay attention to the notation above!

DiD: Setup

Define **potential outcomes** $Y_{it}(z)$ as:

- $Y_{it}(0)$: potential outcome for i in period t when untreated
- $Y_{it}(1)$: potential outcome for i in period t when treated

Note: Pay attention to the notation above!

Individual causal effect for unit i at time t is:

$$\tau_{it} = Y_{it}(1) - Y_{it}(0)$$

DiD: Setup

Define **potential outcomes** $Y_{it}(z)$ as:

- $Y_{it}(0)$: potential outcome for i in period t when untreated
- $Y_{it}(1)$: potential outcome for i in period t when treated

Note: Pay attention to the notation above!

Individual causal effect for unit i at time t is:

$$\tau_{it} = Y_{it}(1) - Y_{it}(0)$$

Observed outcomes Y_{it} are realized as:

$$Y_{it} = Y_{it}(0)(1 - Z_{it}) + Y_{it}(1)Z_{it}$$

DiD: Identification Challenge

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

DiD: Identification Challenge

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 1]$	$\mathbb{E}[Y_{i1}(1) \mid G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 0]$	$\mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

DiD: Identification Challenge

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 1]$	$\mathbb{E}[Y_{i1}(1) \mid G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 0]$	$\mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

Problem: Missing potential outcome $\mathbb{E}[Y_{i1}(0) \mid G_i = 1]$

What would the average post-period outcome for the treated group have been in the absence of treatment?

DiD: Possible Comparisons

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 1]$	$\mathbb{E}[Y_{i1}(1) \mid G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 0]$	$\mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

Comparison: Treated vs. Control, in Post-Period

- ➜ Use $\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 1]$ to estimate ATT
- ➜ Assumes: $\mathbb{E}[Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i0}(0) \mid G_i = 1]$

Read: No change in average potential outcome over time

DiD: Possible Comparisons

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 1]$	$\mathbb{E}[Y_{i1}(1) \mid G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 0]$	$\mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

Comparison: Treated vs. Control, in Post-Period

- ➜ Use $\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]$ to estimate ATT
- ➜ Assumes: $\mathbb{E}[Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

Read: Mean ignorability of treatment assignment

DiD: Possible Comparisons

Estimand: ATT in the post-treatment period

$$ATT = \mathbb{E}[Y_{i1}(1) - Y_{i1}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(1) \mid G_i = 1] - \mathbb{E}[Y_{i1}(0) \mid G_i = 1]$$

Observed quantities:

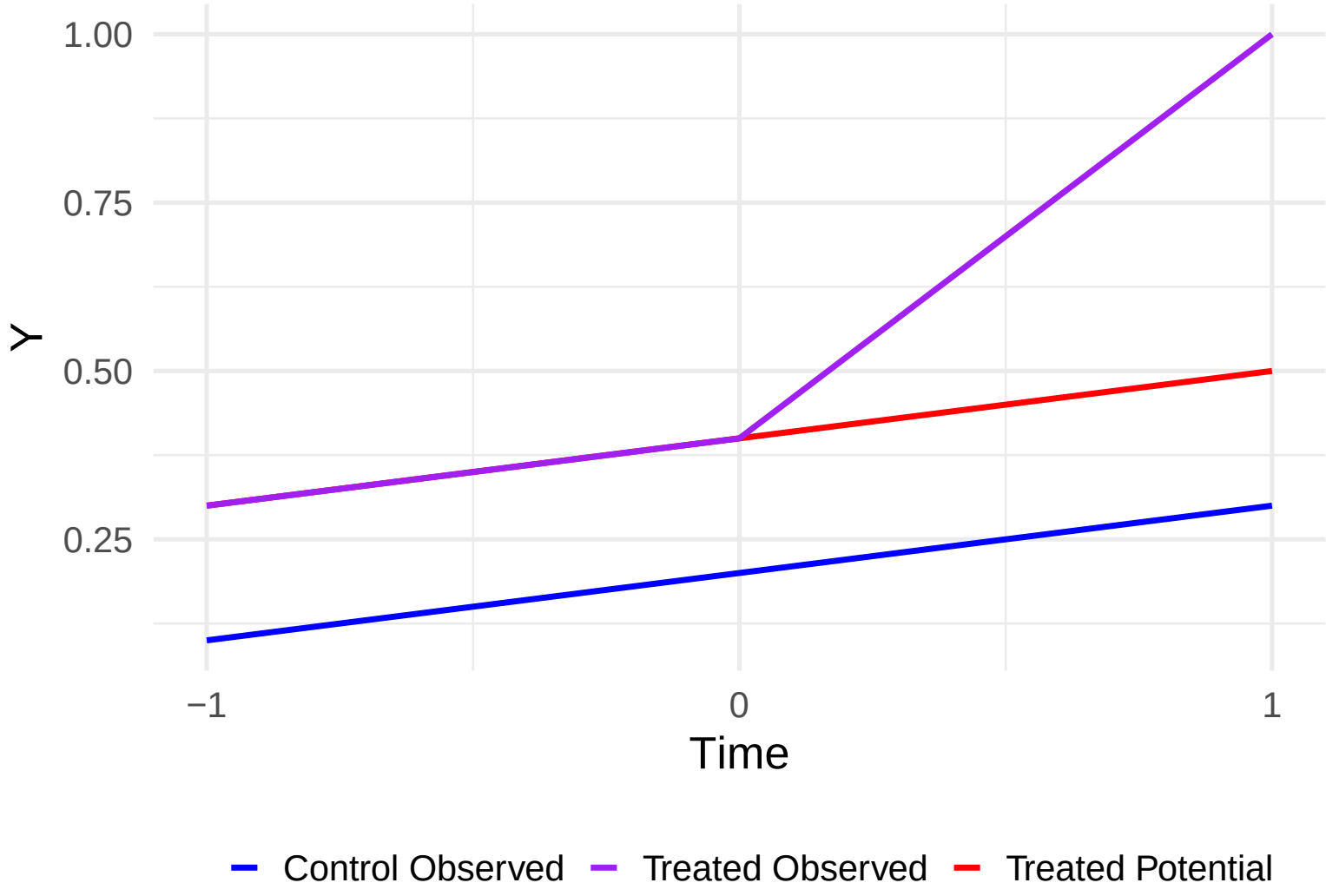
	Pre-Period ($t = 0$)	Post-Period ($t = 1$)
Treatment Group ($G_i = 1$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 1]$	$\mathbb{E}[Y_{i1}(1) \mid G_i = 1]$
Control Group ($G_i = 0$)	$\mathbb{E}[Y_{i0}(0) \mid G_i = 0]$	$\mathbb{E}[Y_{i1}(0) \mid G_i = 0]$

Comparison: Treated vs. Control, in Post-Period

- ➜ Use: $[\mathbb{E}[Y_{i1} \mid G_i = 1] - \mathbb{E}[Y_{i1} \mid G_i = 0]] - [\mathbb{E}[Y_{i0} \mid G_i = 1] - \mathbb{E}[Y_{i0} \mid G_i = 0]]$
- ➜ Assumes: $\mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid G_i = 1] = \mathbb{E}[Y_{i1}(0) - Y_{i0}(0) \mid G_i = 0]$

Read: **Parallel trends**

DiD: Visual Representation of Parallel Trends



DiD: Data Requirements

Consider two data structures: **panel** or **repeated cross-sectional**.

First, let's focus on **panel data**.

A theoretical data structure:

Unit	Time	Y_{it}	G_i	Z_{it}	X_{it}
1	0	$y_{1,0}$	g_1	$z_{1,0}$	$x_{1,0}$
1	1	$y_{1,1}$	g_1	$z_{1,1}$	$x_{1,1}$
2	0	$y_{2,0}$	g_2	$z_{2,0}$	$x_{2,0}$
2	1	$y_{2,1}$	g_2	$z_{2,1}$	$x_{2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	$y_{n,0}$	g_n	$z_{n,0}$	$x_{n,0}$
n	1	$y_{n,1}$	g_n	$z_{n,1}$	$x_{n,1}$

DiD: Data Requirements

Consider two data structures: **panel** or **repeated cross-sectional**.

First, let's focus on **panel data**.

A particular realisation might be:

Unit	Time	Y_{it}	G_i	Z_{it}	X_{it}
1	0	$y_{1,0}$	1	0	$x_{1,0}$
1	1	$y_{1,1}$	1	1	$x_{1,1}$
2	0	$y_{2,0}$	0	0	$x_{2,0}$
2	1	$y_{2,1}$	0	0	$x_{2,1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	0	$y_{n,0}$	1	0	$x_{n,0}$
n	1	$y_{n,1}$	1	1	$x_{n,1}$

DiD: Data Requirements

Now consider the data structure for **repeated cross-sections**.

A particular realisation might be:

Unit	Time	Y_i	G_i	Z_i	X_i
1	0	y_1	g_1	$z_{1,0}$	x_1
2	1	y_2	g_2	$z_{2,1}$	x_2
3	0	y_3	g_3	$z_{3,0}$	x_3
4	1	y_4	g_4	$z_{4,1}$	x_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	0	y_{n-1}	g_{n-1}	z_{n-1}	x_{n-1}
n	1	y_n	g_n	z_n	x_n

DiD: Data Requirements

Now consider the data structure for **repeated cross-sections**.

A particular realisation might be:

Unit	Time	Y_i	G_i	Z_i	X_i
1	0	y_1	1	0	x_1
2	1	y_2	1	1	x_2
3	0	y_3	0	0	x_3
4	1	y_4	0	0	x_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n-1$	0	y_{n-1}	1	0	x_{n-1}
n	1	y_n	1	1	x_n

DiD: Estimation (Repeated Cross-Sections)

Because G_i and T_i are both binary, the same estimator can be calculated via **regression**:

$$\hat{Y}_i = \hat{\mu} + \hat{\gamma} G_i + \hat{\delta} T_i + \hat{\tau} G_i T_i$$

where $\hat{\mu}$, $\hat{\gamma}$, $\hat{\delta}$ and $\hat{\tau}$ are estimated with OLS regression.

It's easy to show that $\hat{\tau} = A\hat{T}T$:

	After ($T_i = 1$)	Before ($T_i = 0$)	After $-$ Before
Treated ($G_i = 1$)	$\hat{\mu} + \hat{\gamma} + \hat{\delta} + \hat{\tau}$	$\hat{\mu} + \hat{\gamma}$	$\hat{\delta} + \hat{\tau}$
Control ($G_i = 0$)	$\hat{\mu} + \hat{\delta}$	$\hat{\mu}$	$\hat{\delta}$
Treated $-$ Control	$\hat{\gamma} + \hat{\tau}$	$\hat{\gamma}$	$\hat{\tau}$

DiD: Estimation (Panel Data)

For panel data, consider an **additive linear model** for potential outcomes:

$$Y_{it}(z) = \alpha_i + \gamma t + \tau z + \epsilon_{it}$$

where α_i is a **time-invariant unobserved parameter** for unit i.

The **first-differenced regression** of $\Delta Y_i = Y_{i1} - Y_{i0}$ on G_i can unbiasedly estimate $ATT = ATE$.

Notice that panel data allow for *unit-level* unobserved confounding beyond *group-level* unobserved confounding, but it must be **additive** and **time-invariant**.

DiD: Estimation (Panel Data)

What if we have many periods?

Standard for many years was the 'two-way fixed effects' regression:

$$\hat{Y}_{it} = \hat{\alpha}_i + \hat{\gamma}_t + \hat{\tau} Z_{it}$$

Where α_i is a unit fixed effect, γ_t is a time fixed effect, and Z_{it} gives time-varying treatment status.

As before, $\hat{\tau}$ is the DiD estimator of *ATT* (and *ATE* if our DGP is right).

DiD: Good Practice

1. The key assumption is parallel trends – we cannot test this directly. (Why?). Instead, test parallel pre-trends.

2. For canonical DiD, use `lm()` or `lm_robust()` for cross-sectional, or `fixest::feols()` for panel.

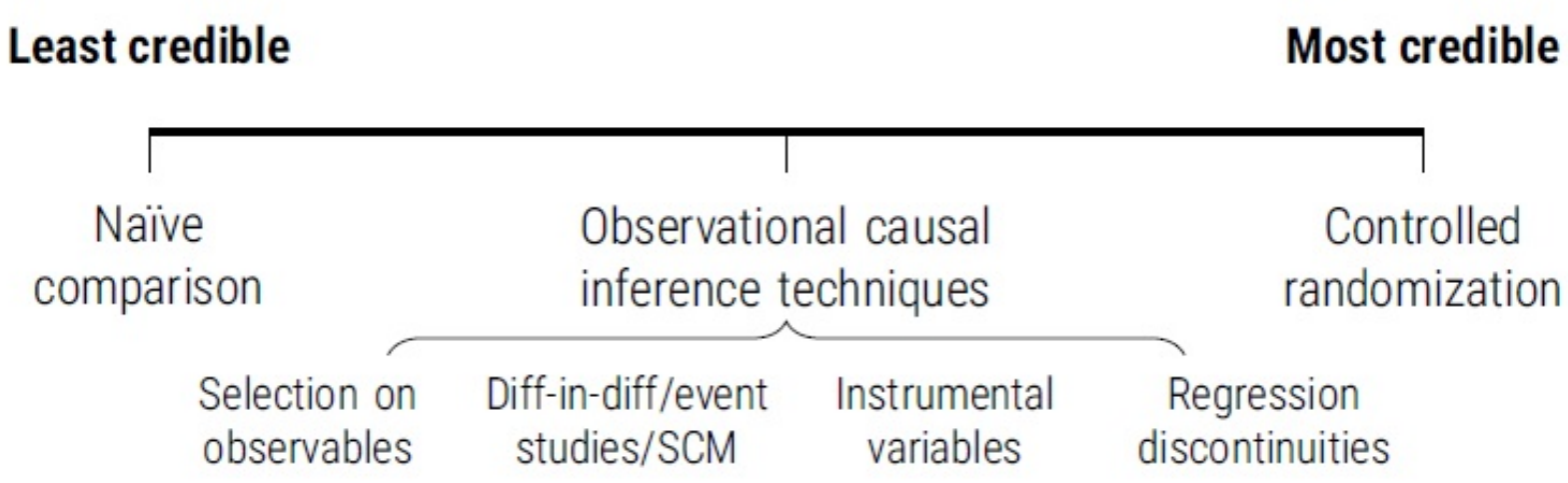
3. We have covered 'canonical 2-period difference-in-differences.' What if you have more that two periods, and treatment is staggered?

4. For these cases, look at `{fect}` and `{did}` packages in `R`.

Causal Inference: Tokyo Drift

The Lamp Post Problem

Continuum of 'Credibility'™



The art (and science) of applied causal inference is making defensible **assumptions**. There is no 'magic' to the FPCI, just **assumptions** all the way down.

Assumptions All the Way Down

Wrapping Up

Today we covered:

A trio of ways to make causal claims in the absence of randomized D:

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

Tomorrow:

- ➜ Intro to machine learning (ML): predicting 'missing values'
- ➜ Performance metrics for ML
- ➜ Doing prediction with regression and naive Bayes