Some review and extra stuff

Francisco Villamil

Applied Quantitative Methods II

MA in Social Sciences, Spring 2026

Today's second half

- The course has been about **estimation**: how to fit models

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question
  - → Research design and identification

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question
  - → Research design and identification
  - → Causal inference methods beyond FE/DiD

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question
  - → Research design and identification
  - → Causal inference methods beyond FE/DiD
  - → Defending your results: robustness and validity

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question
  - → Research design and identification
  - → Causal inference methods beyond FE/DiD
  - → Defending your results: robustness and validity
  - → Common pitfalls when applying all of this

Today's second half

- The course has been about **estimation**: how to fit models
- But fitting a model is rarely the hard part of empirical work
- Today: what you actually need to answer a research question
  - → Research design and identification
  - → Causal inference methods beyond FE/DiD
  - → Defending your results: robustness and validity
  - → Common pitfalls when applying all of this
- Plus a pointer to two advanced topics (not covered today)

Roadmap

From estimation to identification

Causal inference toolkit

Defending a causal claim

Choosing the right method & pitfalls

What you have learned

- Estimation tools for many kinds of outcomes

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure
  - → Fixed effects, two-way FE, DiD (including staggered)

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure
  - → Fixed effects, two-way FE, DiD (including staggered)
  - → Spatial models

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure
  - → Fixed effects, two-way FE, DiD (including staggered)
  - → Spatial models
- How to interpret and present results

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure
  - → Fixed effects, two-way FE, DiD (including staggered)
  - → Spatial models
- How to interpret and present results
  - → Predicted values, marginal effects, coefficient plots

What you have learned

- Estimation tools for many kinds of outcomes
  - → OLS, logit, count / ordinal / duration models
- Designs that exploit data structure
  - → Fixed effects, two-way FE, DiD (including staggered)
  - → Spatial models
- How to interpret and present results
  - → Predicted values, marginal effects, coefficient plots
- Plus computing practices to make all of this reproducible

Two different questions

Estimation: given a model, what are the parameters?

Identification: do those parameters mean what I think they mean?

Two different questions

Estimation: given a model, what are the parameters?

Identification: do those parameters mean what I think they mean?

Your R output answers the first.

Only your research design can answer the second.

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise
- For unit i with treatment D:

$$\tau_i = Y_i(1) - Y_i(0)$$

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise
- For unit i with treatment D:

$$\tau_i = Y_i(1) - Y_i(0)$$

- We only ever observe **one** of these two worlds

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise
- For unit i with treatment D:

$$\tau_i = Y_i(1) - Y_i(0)$$

- We only ever observe **one** of these two worlds
  - $\rightarrow$ If i was treated, $Y_i(1)$ is observed and $Y_i(0)$ is missing

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise
- For unit i with treatment D:

$$\tau_i = Y_i(1) - Y_i(0)$$

- We only ever observe **one** of these two worlds
  - $\rightarrow$ If i was treated, $Y_i(1)$ is observed and $Y_i(0)$ is missing
  - $\rightarrow$ If i was not treated, the opposite

The fundamental problem of causal inference

- A **causal effect** is the difference between what *did* happen and what *would have* happened otherwise
- For unit i with treatment D:

$$\tau_i = Y_i(1) - Y_i(0)$$

- We only ever observe **one** of these two worlds
  - → If i was treated, $Y_i(1)$ is observed and $Y_i(0)$ is missing
  - → If i was not treated, the opposite
- Causal inference = principled way to estimate the missing counterfactual

What we can estimate on average

- We can't recover τ_i for individuals
- But we can target averages:
  - → **ATE** — average over everyone: $E[Y(1) - Y(0)]$
  - → **ATT** — average over the treated: $E[Y(1) - Y(0) \mid D = 1]$
  - → **LATE** — average over those who *respond* to a treatment (IV, RDD)
- Different methods estimate different quantities
  - → When you compare two papers: are they even targeting the same thing?

Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place

Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects

Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects
- Three structures you must recognize:

Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects
- Three structures you must recognize:
  - → **Confounder** $D \leftarrow Z \rightarrow Y$

Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects
- Three structures you must recognize:
  - → **Confounder** $D \leftarrow Z \rightarrow Y$
  - → **Mediator** $D \rightarrow M \rightarrow Y$

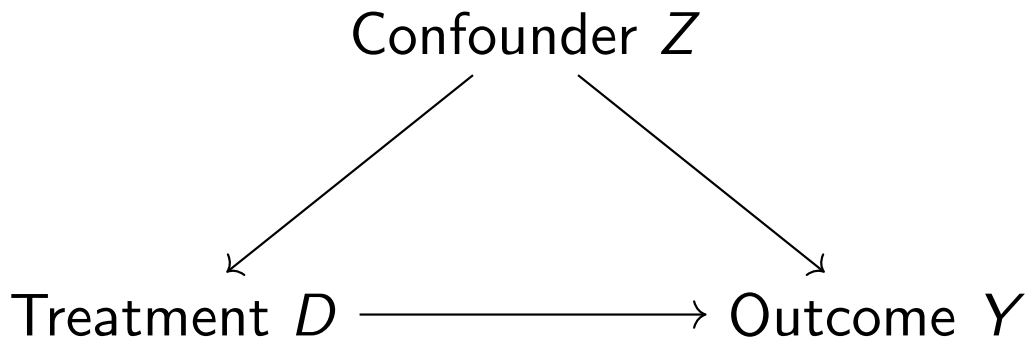
Causal models and DAGs

- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects
- Three structures you must recognize:
  - → **Confounder** $D \leftarrow Z \rightarrow Y$
  - → **Mediator** $D \rightarrow M \rightarrow Y$
  - → **Collider** $D \rightarrow C \leftarrow Y$

Causal models and DAGs

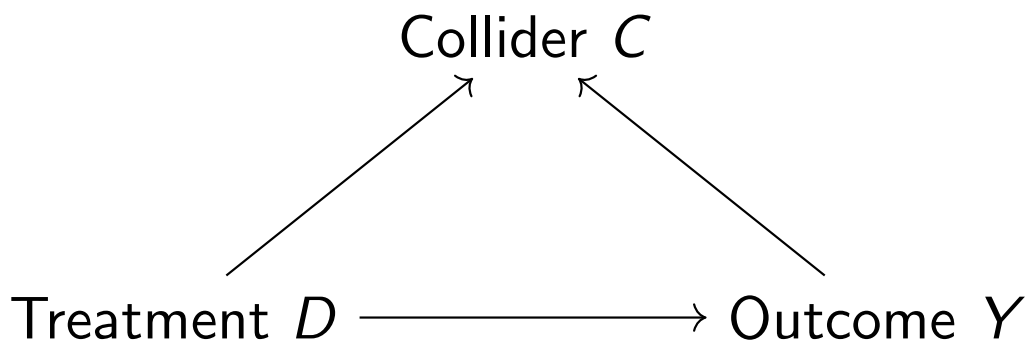
- To approximate the counterfactual we need a model of *why* units are treated in the first place
- Directed Acyclic Graphs (DAGs): nodes are variables, arrows are direct causal effects
- Three structures you must recognize:
  - → **Confounder** $D \leftarrow Z \rightarrow Y$
  - → **Mediator** $D \rightarrow M \rightarrow Y$
  - → **Collider** $D \rightarrow C \leftarrow Y$
- What you control for — and what you do **not** — depends on which structure you are looking at

Confounding: the standard case



- Z affects both D and Y: a back-door path
- Un-adjusted comparison of treated vs. untreated mixes two effects
- **Solution:** condition on Z (or use a design that blocks the back door)

Colliders: do not control!



- C is caused *by* both D and Y
- Conditioning on C creates a spurious association between D and Y
- Classic: "smart people lack social skills" (if you condition on being hired)

Post-treatment bias

- Special case of collider bias that is very easy to miss

Post-treatment bias

- Special case of collider bias that is very easy to miss
- A **post-treatment** variable is one that is itself affected by the treatment

Post-treatment bias

- Special case of collider bias that is very easy to miss
- A **post-treatment** variable is one that is itself affected by the treatment
- Controlling for it:

Post-treatment bias

- Special case of collider bias that is very easy to miss
- A **post-treatment** variable is one that is itself affected by the treatment
- Controlling for it:
  - $\rightarrow$ blocks part of the mechanism (if it is a mediator), **or**

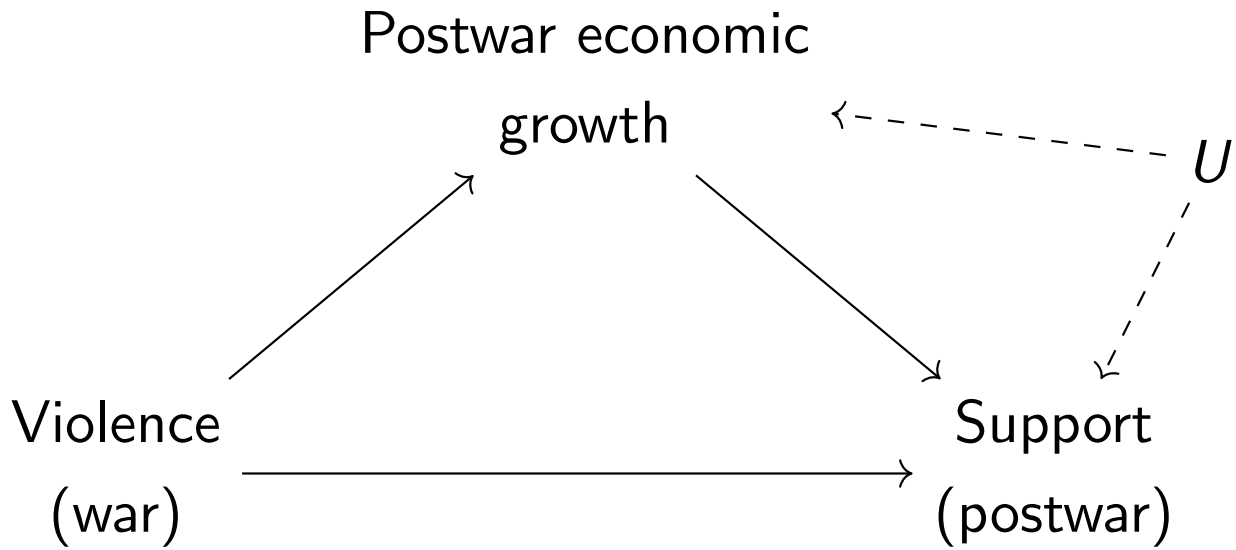
Post-treatment bias

- Special case of collider bias that is very easy to miss
- A **post-treatment** variable is one that is itself affected by the treatment
- Controlling for it:
  - → blocks part of the mechanism (if it is a mediator), **or**
  - → opens a collider path (if it is also affected by other unobservables)

Post-treatment bias

- Special case of collider bias that is very easy to miss
- A **post-treatment** variable is one that is itself affected by the treatment
- Controlling for it:
  - → blocks part of the mechanism (if it is a mediator), **or**
  - → opens a collider path (if it is also affected by other unobservables)
- Rule of thumb: if in doubt about whether a variable is pre- or post-treatment, leave it out

Post-treatment bias: example



- Does wartime violence affect postwar political support?
- "Let's control for postwar economic growth, it also affects support"
- But growth is **caused by** the treatment *and* shares unobservables with the outcome — controlling for it opens a back door

So what does your regression assume?

Every regression coefficient you report implicitly assumes:

1. You have **closed all back doors** (no omitted confounders)
2. You have **not opened new ones** (no controlling for colliders or mediators)
3. The functional form is roughly right (linearity, additivity)

So what does your regression assume?

Every regression coefficient you report implicitly assumes:

1. You have **closed all back doors** (no omitted confounders)
2. You have **not opened new ones** (no controlling for colliders or mediators)
3. The functional form is roughly right (linearity, additivity)

(1) and (2) are **design** assumptions, not something you can test.

Roadmap

From estimation to identification

Causal inference toolkit

Defending a causal claim

Choosing the right method & pitfalls

What you already have

Two of the four big causal-inference designs are already in your toolkit:

- **Fixed effects** — control for all time-invariant unobservables at some level (unit, region, year. . .)
- **Difference-in-differences** — use a never-treated comparison group plus pre/post variation; identifies the ATT under parallel trends

Two more that you should at least recognize:

- **Regression discontinuity design (RDD)**
- **Instrumental variables (IV)**

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

- β_3 is the DiD estimate

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

- β_3 is the DiD estimate
- In practice: two-way fixed effects absorb Treat_i and Post_t

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

- β_3 is the DiD estimate

- In practice: two-way fixed effects absorb Treat_i and Post_t

```
feols(y ~ treat_post | unit + time, data = d)
```

Estimating DiD

- DiD is not a special estimator — it's just OLS on a 2×2 interaction

$$Y_{it} = \alpha + \beta_1 \text{Treat}_i + \beta_2 \text{Post}_t + \beta_3 (\text{Treat}_i \times \text{Post}_t) + \varepsilon_{it}$$

- β_3 is the DiD estimate
- In practice: two-way fixed effects absorb Treat_i and Post_t

```
feols(y ~ treat_post | unit + time, data = d)
```

- For staggered adoption: use modern estimators (Callaway–Sant'Anna, Sun–Abraham) — see Session 6

Regression discontinuity design

- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff

Regression discontinuity design

- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff
  - $\rightarrow$ Incumbency advantage: winners of close elections

Regression discontinuity design

- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff
  - → Incumbency advantage: winners of close elections
  - → Scholarships: eligibility based on test score cutoff

Regression discontinuity design

- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff
  - → Incumbency advantage: winners of close elections
  - → Scholarships: eligibility based on test score cutoff
  - → Conscription: draft lottery numbers below a threshold

Regression discontinuity design

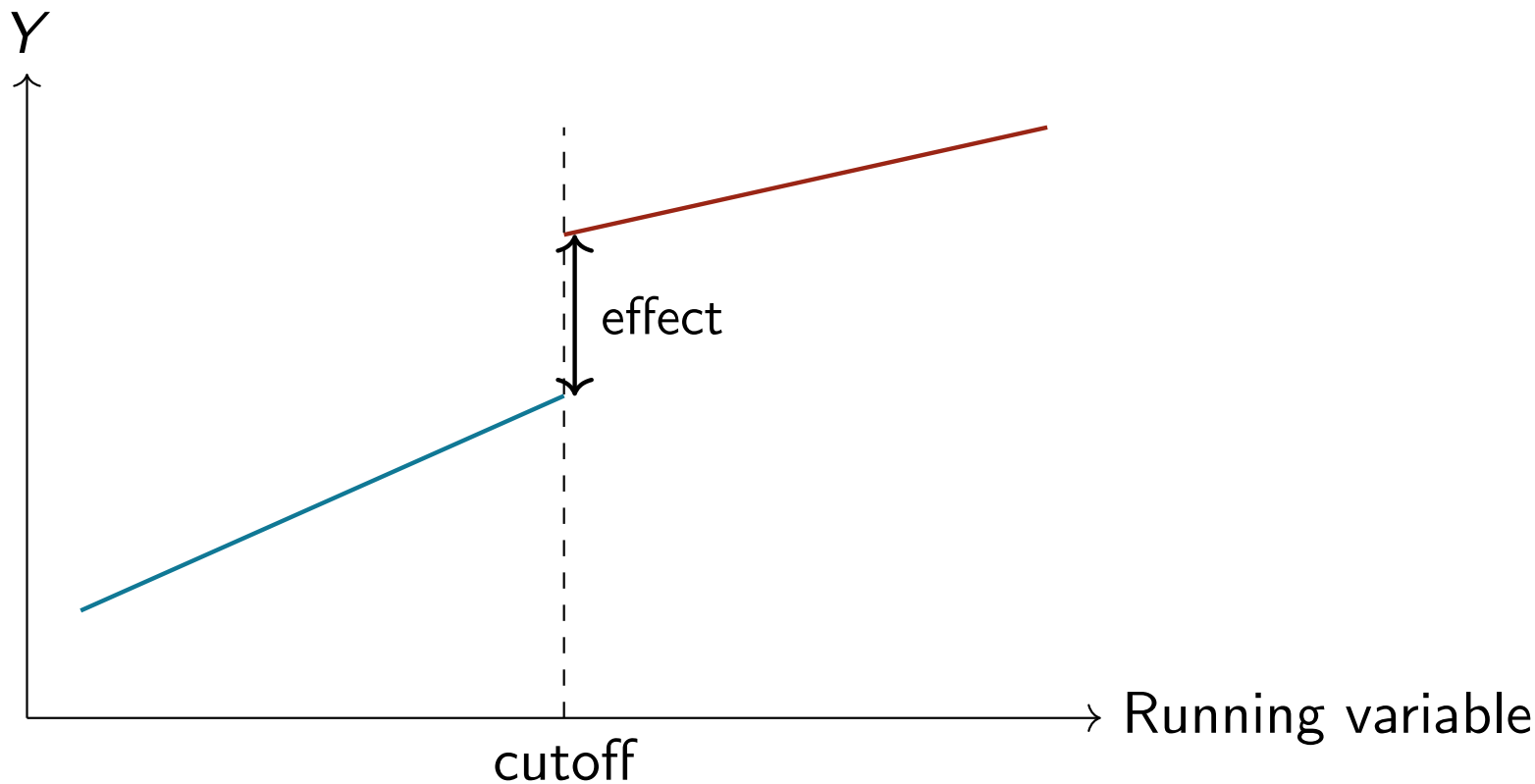
- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff
  - → Incumbency advantage: winners of close elections
  - → Scholarships: eligibility based on test score cutoff
  - → Conscription: draft lottery numbers below a threshold

- **Idea:** units just above and just below the cutoff should be similar in everything *except* the treatment

Regression discontinuity design

- **Setup:** treatment assignment depends on whether a *running variable* crosses a cutoff
  - → Incumbency advantage: winners of close elections
  - → Scholarships: eligibility based on test score cutoff
  - → Conscription: draft lottery numbers below a threshold
- **Idea:** units just above and just below the cutoff should be similar in everything *except* the treatment
- Estimates a LATE: effect for units near the threshold

RDD visually



- The jump at the cutoff is the treatment effect
- The trend on each side controls for whatever else varies smoothly

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

- τ is the RDD estimate (the jump at the cutoff)

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

- τ is the RDD estimate (the jump at the cutoff)
- Key choice: bandwidth — how wide a window around the cutoff

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

- τ is the RDD estimate (the jump at the cutoff)
- Key choice: bandwidth — how wide a window around the cutoff
  - → Data-driven: Calonico, Cattaneo & Titiunik

Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

- τ is the RDD estimate (the jump at the cutoff)
- Key choice: bandwidth — how wide a window around the cutoff
  - → Data-driven: Calonico, Cattaneo & Titiunik
- R: `rdrobust::rdrobust(y, x, c = 0)`

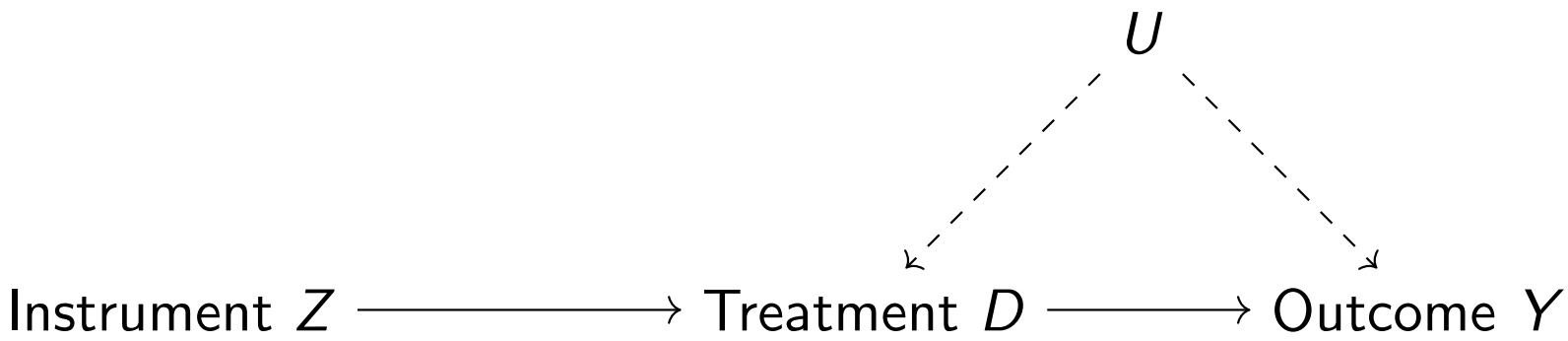
Estimating RDD

- Center the running variable at the cutoff: $\tilde{X}_i = X_i - c$
- Define treatment: $D_i = \mathbf{1}(\tilde{X}_i \geq 0)$
- Run a local linear regression with slopes allowed to differ on each side:

$$Y_i = \alpha + \tau D_i + \beta_1 \tilde{X}_i + \beta_2 (D_i \times \tilde{X}_i) + \varepsilon_i$$

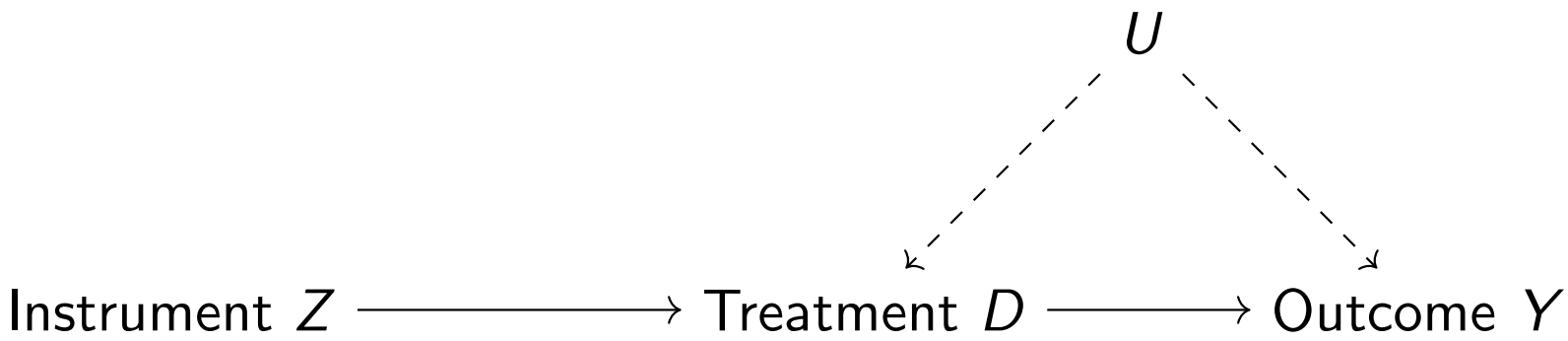
- τ is the RDD estimate (the jump at the cutoff)
- Key choice: bandwidth — how wide a window around the cutoff
  - → Data-driven: Calonico, Cattaneo & Titiunik
- R: `rdrobust::rdrobust(y, x, c = 0)`
  - → `c` is the cutoff value of the running variable `x`

Instrumental variables



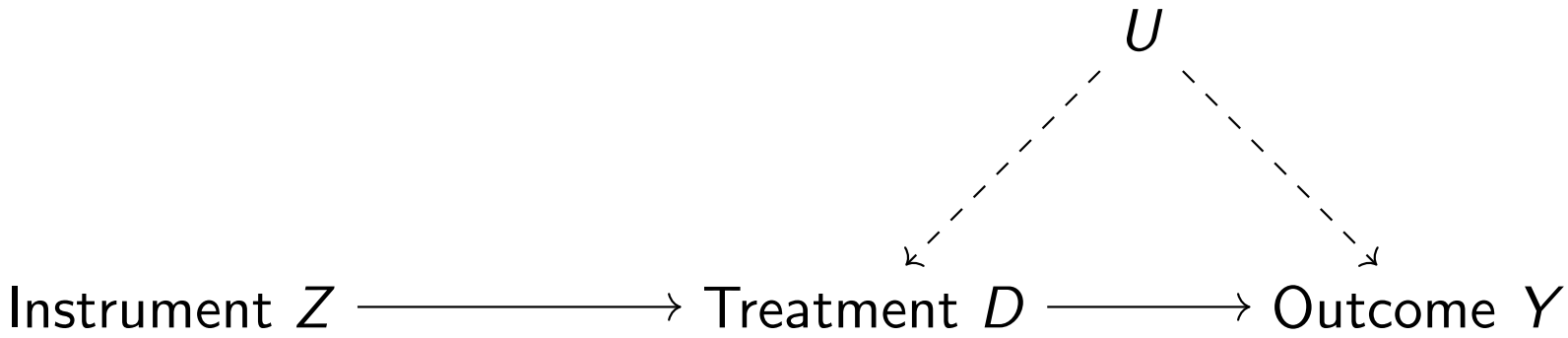
- Find a variable Z that moves D but has no other path to Y

Instrumental variables



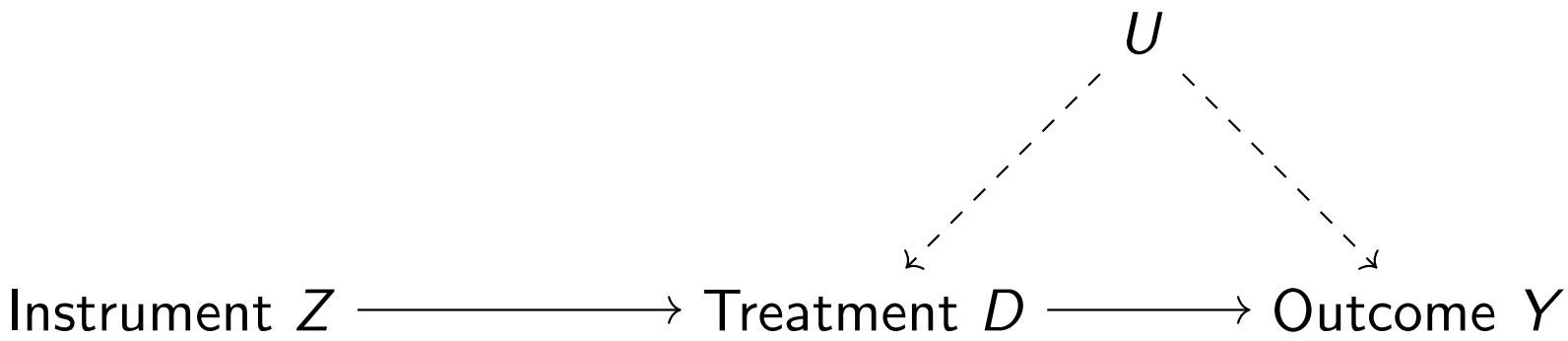
- Find a variable Z that moves D but has no other path to Y
- Use only the variation in D explained by Z

Instrumental variables



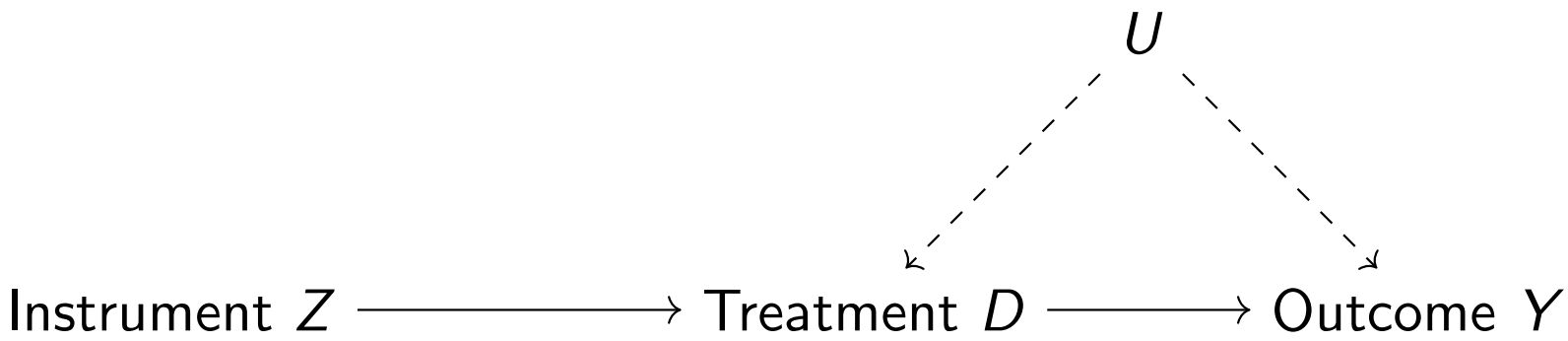
- Find a variable Z that moves D but has no other path to Y
- Use only the variation in D explained by Z
- Two assumptions:

Instrumental variables



- Find a variable Z that moves D but has no other path to Y
- Use only the variation in D explained by Z
- Two assumptions:
  - → **Relevance**: Z actually affects D (testable)

Instrumental variables



- Find a variable Z that moves D but has no other path to Y
- Use only the variation in D explained by Z
- Two assumptions:
  - → **Relevance**: Z actually affects D (testable)
  - → **Exclusion restriction**: Z affects Y only through D (not testable)

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

$$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

$$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

  $$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

  $$Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i$$

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

  $$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

  $$Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i$$

- β is the IV estimate: a LATE for units whose D responds to Z

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

  $$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

  $$Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i$$

- β is the IV estimate: a LATE for units whose D responds to Z

- Always report the first stage, and an F-statistic on π_1

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

  $$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

  $$Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i$$

- β is the IV estimate: a LATE for units whose D responds to Z

- Always report the first stage, and an F-statistic on π_1
  - $\rightarrow$ Rule of thumb: $F < 10$ signals a weak instrument

Estimating IV: two-stage least squares

- **First stage:** regress the treatment on the instrument

  $$D_i = \pi_0 + \pi_1 Z_i + u_i \quad \Rightarrow \quad \hat{D}_i$$

- **Second stage:** regress the outcome on the *predicted* treatment

  $$Y_i = \alpha + \beta \hat{D}_i + \varepsilon_i$$

- β is the IV estimate: a LATE for units whose D responds to Z

- Always report the first stage, and an F-statistic on π_1

  $\rightarrow$ Rule of thumb: $F < 10$ signals a weak instrument

- R: `feols(y ~ 1 | d ~ z)` or `AER::ivreg()`

The four design templates

Design	**Exogenous variation**	**Key assumption**
FE	Within-unit differences	No time-varying confounders
DiD	Treated vs. comparison, pre/post	Parallel trends
RDD	Jump at cutoff	No manipulation at cutoff
IV	Instrument $\rightarrow$ treatment	Exclusion restriction

- You do not pick a design, *the data* pick one for you
- What source of exogenous variation does the world hand you?

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step
- **Idea:** pair each treated unit with one or more untreated units that look similar on observed covariates

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step
- **Idea:** pair each treated unit with one or more untreated units that look similar on observed covariates
- Compare only within matched pairs

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step
- **Idea:** pair each treated unit with one or more untreated units that look similar on observed covariates
- Compare only within matched pairs
- What it fixes: imbalance in observed covariates and reliance on linear-in-parameters extrapolation

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step
- **Idea:** pair each treated unit with one or more untreated units that look similar on observed covariates
- Compare only within matched pairs
- What it fixes: imbalance in observed covariates and reliance on linear-in-parameters extrapolation
- What it does **not** fix: omitted confounders, selection on unobservables

Matching: complement, not substitute

- Not a design in the same sense — but a useful preprocessing step
- **Idea:** pair each treated unit with one or more untreated units that look similar on observed covariates
- Compare only within matched pairs
- What it fixes: imbalance in observed covariates and reliance on linear-in-parameters extrapolation
- What it does **not** fix: omitted confounders, selection on unobservables
- Often used *with* regression or DiD, not instead of

Roadmap

From estimation to identification

Causal inference toolkit

Defending a causal claim

Choosing the right method & pitfalls

Your main result is the beginning

- You have a point estimate, a standard error, a figure
- A skeptical reader should be asking:
  - → What if you had specified it differently?
  - → What if your identifying assumption is wrong?
  - → What if you are measuring the wrong thing?
  - → Would the result generalize elsewhere?
- Your job is to answer those questions **before they do**

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales
- Measurement of D: dichotomize vs. continuous, source

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales
- Measurement of D: dichotomize vs. continuous, source
- Sample: drop influential units, drop specific sub-groups

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales
- Measurement of D: dichotomize vs. continuous, source
- Sample: drop influential units, drop specific sub-groups
- Model: add/remove controls, change FE structure, change SEs

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales
- Measurement of D: dichotomize vs. continuous, source
- Sample: drop influential units, drop specific sub-groups
- Model: add/remove controls, change FE structure, change SEs
- Functional form: log, quadratic, non-parametric

Robustness: alternative specifications

- Change one thing at a time, show it does not matter much
- Measurement of Y: alternative indicators, different scales
- Measurement of D: dichotomize vs. continuous, source
- Sample: drop influential units, drop specific sub-groups
- Model: add/remove controls, change FE structure, change SEs
- Functional form: log, quadratic, non-parametric
- Display as a coefficient plot across specifications, not a 10-column table no one reads

Placebo tests

- Look for an effect *where there should be none*

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect
  - → **Placebo population**: a subgroup not exposed to D

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect
  - → **Placebo population**: a subgroup not exposed to D
- Example — Peisakhin & Rozenas (2018), Russian TV in Ukraine:

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect
  - → **Placebo population**: a subgroup not exposed to D
- Example — Peisakhin & Rozenas (2018), Russian TV in Ukraine:
  - Outcome: signal strength should not predict % Russian speakers

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect
  - → **Placebo population**: a subgroup not exposed to D
- Example — Peisakhin & Rozenas (2018), Russian TV in Ukraine:
  - Outcome: signal strength should not predict % Russian speakers
  - Treatment: Russian *sports* TV signal should not predict voting

Placebo tests

- Look for an effect *where there should be none*
- If you find it, something is wrong with your design
- Three flavors (Eggers, Tuñón & Dafoe 2023):
  - → **Placebo outcome**: different Y that the treatment should not affect
  - → **Placebo treatment**: a fake D that should have no effect
  - → **Placebo population**: a subgroup not exposed to D
- Example — Peisakhin & Rozenas (2018), Russian TV in Ukraine:
  - Outcome: signal strength should not predict % Russian speakers
  - Treatment: Russian *sports* TV signal should not predict voting
  - Population: Ukrainians on satellite TV — signal irrelevant

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel
- If pre-treatment "leads" are significant, parallel trends is suspect

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel
- If pre-treatment "leads" are significant, parallel trends is suspect
- Event-study plots show this visually:

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel
- If pre-treatment "leads" are significant, parallel trends is suspect

- Event-study plots show this visually:
  - $\rightarrow$ Coefficients on event-time dummies

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel
- If pre-treatment "leads" are significant, parallel trends is suspect

- Event-study plots show this visually:
  - → Coefficients on event-time dummies
  - → Pre-period coefficients should be flat and near zero

Placebo tests in DiD

- Most important placebo in a DiD setup: **pre-trends**
- Before the treatment, the two groups should move in parallel
- If pre-treatment "leads" are significant, parallel trends is suspect

- Event-study plots show this visually:
  - → Coefficients on event-time dummies
  - → Pre-period coefficients should be flat and near zero
  - → Post-period coefficients show the dynamic effect

Additional implications

- If your theory is correct, what *else* should be true in the data?

Additional implications

- If your theory is correct, what *else* should be true in the data?
- **Mechanism checks**: intermediate variables your theory predicts

Additional implications

- If your theory is correct, what *else* should be true in the data?
- **Mechanism checks**: intermediate variables your theory predicts
- **Heterogeneity**: groups where the effect should be larger or smaller

Additional implications

- If your theory is correct, what *else* should be true in the data?
- **Mechanism checks**: intermediate variables your theory predicts
- **Heterogeneity**: groups where the effect should be larger or smaller
- **Rule-out tests**: observable patterns that competing explanations would imply but yours would not

Additional implications

- If your theory is correct, what *else* should be true in the data?
- **Mechanism checks**: intermediate variables your theory predicts
- **Heterogeneity**: groups where the effect should be larger or smaller
- **Rule-out tests**: observable patterns that competing explanations would imply but yours would not
- A paper that only shows the main effect is weaker than one that shows the main effect **plus** two or three such auxiliary findings

Measurement and construct validity

- Are you measuring what you think you are measuring?

Measurement and construct validity

- Are you measuring what you think you are measuring?
- **Treatment**: when you say $D = 1$, what is actually being delivered?

Measurement and construct validity

- Are you measuring what you think you are measuring?
- **Treatment**: when you say $D = 1$, what is actually being delivered?

  $\rightarrow$ Bundled treatments (media exposure = content + format + novelty...)

Measurement and construct validity

- Are you measuring what you think you are measuring?
- **Treatment**: when you say $D = 1$, what is actually being delivered?

  $\rightarrow$ Bundled treatments (media exposure = content + format + novelty...)

- **Outcome**: does the indicator really capture the concept?

Measurement and construct validity

- Are you measuring what you think you are measuring?
- **Treatment**: when you say $D = 1$, what is actually being delivered?
  - → Bundled treatments (media exposure = content + format + novelty...)
- **Outcome**: does the indicator really capture the concept?
  - → Polarization = partisan thermometer gap? policy-view distance? affective feelings?

Measurement and construct validity

- Are you measuring what you think you are measuring?
- **Treatment**: when you say $D = 1$, what is actually being delivered?
  - → Bundled treatments (media exposure = content + format + novelty...)
- **Outcome**: does the indicator really capture the concept?
  - → Polarization = partisan thermometer gap? policy-view distance? affective feelings?
- Identification can be perfect and the claim still be wrong if the measurement doesn't line up with the theory

External validity

- Under what *conditions* does the effect hold?

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample
  - → **Treatments**: variants of the intervention

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample
  - → **Treatments**: variants of the intervention
  - → **Outcomes**: related but different Ys

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample
  - → **Treatments**: variants of the intervention
  - → **Outcomes**: related but different Ys
  - → **Contexts**: other settings, places, times

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample
  - → **Treatments**: variants of the intervention
  - → **Outcomes**: related but different Ys
  - → **Contexts**: other settings, places, times

- A clean estimate in one sample does not automatically generalize

External validity

- Under what *conditions* does the effect hold?
- Four dimensions (Egami & Hartman 2023):
  - → **Populations**: units beyond your sample
  - → **Treatments**: variants of the intervention
  - → **Outcomes**: related but different Ys
  - → **Contexts**: other settings, places, times

- A clean estimate in one sample does not automatically generalize
- **Temporal validity** (Munger 2023): does it still hold 10 years from now?

Roadmap

From estimation to identification

Causal inference toolkit

Defending a causal claim

Choosing the right method & pitfalls

Starting point: your question type

Goal	**Tools**
Describe a pattern	Summary stats, visualization, simple regression
Predict Y from X	Regression, ML (cross-validated)
Explain Y causally	Design-based methods: FE, DiD, RDD, IV, experiments
Generalize a mechanism	Theory + replication across contexts

- Most confusion in applied work comes from mixing these up
  - $\rightarrow$ e.g. treating predictive coefficients as causal

Starting point: your outcome

Outcome type	**Method**
Continuous	OLS (always a good first pass)
Binary	Logit, probit (or LPM for interpretability)
Count (non-negative integer)	Poisson, negative binomial
Ordered categories	Ordered logit / probit
Time-to-event	Cox, Kaplan–Meier, parametric survival
Panel data	+ fixed effects, + clustered SEs
Spatial data	+ spatial weights, + Moran's I

- This is a cheat-sheet for your final essay — keep it

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include
  - → How to code the treatment

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include
  - → How to code the treatment
  - → Which subsample to focus on

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include
  - → How to code the treatment
  - → Which subsample to focus on
  - → How to handle outliers

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include
  - → How to code the treatment
  - → Which subsample to focus on
  - → How to handle outliers
- If any of those choices were made *after looking at the results*, your p-values are too small

Pitfall 1: p-hacking and forking paths

- Garden of forking paths (Gelman & Loken 2013)
- Every analysis involves many small choices:
  - → Which controls to include
  - → How to code the treatment
  - → Which subsample to focus on
  - → How to handle outliers
- If any of those choices were made *after looking at the results*, your p-values are too small
- Mitigations: pre-analysis plans, specification curves, honest reporting of all tried specifications

Pitfall 2: bad controls

- "More controls" is not always better

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - $\rightarrow$ Mediators (blocks the effect you want to measure)

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - → Mediators (blocks the effect you want to measure)
  - → Colliders (opens a spurious path)

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - → Mediators (blocks the effect you want to measure)
  - → Colliders (opens a spurious path)
  - → Post-treatment variables (same problem, often both)

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - → Mediators (blocks the effect you want to measure)
  - → Colliders (opens a spurious path)
  - → Post-treatment variables (same problem, often both)
  - → Proxies for the treatment (collinearity)

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - → Mediators (blocks the effect you want to measure)
  - → Colliders (opens a spurious path)
  - → Post-treatment variables (same problem, often both)
  - → Proxies for the treatment (collinearity)
- Every control needs a justification from your DAG

Pitfall 2: bad controls

- "More controls" is not always better
- **Do not control for**:
  - → Mediators (blocks the effect you want to measure)
  - → Colliders (opens a spurious path)
  - → Post-treatment variables (same problem, often both)
  - → Proxies for the treatment (collinearity)
- Every control needs a justification from your DAG
- Cinelli, Forney & Pearl (2024) "A crash course in good and bad controls"

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms
  - → Same unit observed over time (panel)

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms
  - → Same unit observed over time (panel)
  - → Spatial proximity

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms
  - → Same unit observed over time (panel)
  - → Spatial proximity
  - → Network connections

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms
  - → Same unit observed over time (panel)
  - → Spatial proximity
  - → Network connections
- Consequence: default SEs can be *much* too small

Pitfall 3: ignoring dependence

- Standard errors assume observations are independent
- In social science data they usually are not:
  - → Observations nested within countries, schools, firms
  - → Same unit observed over time (panel)
  - → Spatial proximity
  - → Network connections
- Consequence: default SEs can be *much* too small
- Fixes: cluster-robust SEs, Conley SEs (spatial), network-adjusted SEs

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?
- What method matches the question and the variation?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?
- What method matches the question and the variation?
- What are the **identifying assumptions**, and how plausible are they?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?
- What method matches the question and the variation?
- What are the **identifying assumptions**, and how plausible are they?
- What **alternative explanations** does the design *not* rule out?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?
- What method matches the question and the variation?
- What are the **identifying assumptions**, and how plausible are they?
- What **alternative explanations** does the design *not* rule out?
- What **robustness / placebo** checks can you run?

A checklist for your final essay

- What is the **question** (descriptive, predictive, causal)?
- What is the **unit** of analysis and where does **variation** come from?
- What is the **DAG** behind the relationship?
- What method matches the question and the variation?
- What are the **identifying assumptions**, and how plausible are they?
- What **alternative explanations** does the design *not* rule out?
- What **robustness / placebo** checks can you run?
- How do you **measure** your key concepts, and what's lost in that translation?

Roadmap

Appendix: Multilevel models

Appendix: Text as data

Hierarchical data

- Very common structure in social science:

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries
  - → Respondents within country-years

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries
  - → Respondents within country-years
- Two facts:

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries
  - → Respondents within country-years
- Two facts:
  - → Units within the same group are more alike than units in different groups

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries
  - → Respondents within country-years
- Two facts:
  - → Units within the same group are more alike than units in different groups
  - → Groups vary in how they respond to predictors

Hierarchical data

- Very common structure in social science:
  - → Students within schools within districts
  - → Voters within precincts within countries
  - → Respondents within country-years
- Two facts:
  - → Units within the same group are more alike than units in different groups
  - → Groups vary in how they respond to predictors
- Ordinary OLS ignores both

Three ways to handle groups

- **Complete pooling**: ignore groups entirely, one intercept
  → Wastes the group-level information
- **No pooling**: separate intercept for each group (fixed effects)
  → Fits each group exactly — but noisily for small groups
- **Partial pooling**: each group gets its own intercept, but shrunk toward a global mean
  → Best of both worlds — this is what multilevel models do

Random intercepts

$$y_{ij} = \alpha_j + \beta x_{ij} + \varepsilon_{ij}, \quad \alpha_j \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$$

- Each group j has its own intercept α_j
- But the α_j's are drawn from a common distribution
- Estimate: one global mean + one variance, not 200 separate intercepts
- Versus classic FE: FE *estimates* each α_j; random intercepts *model* them

Random slopes

$$y_{ij} = \alpha_j + \beta_j x_{ij} + \varepsilon_{ij}$$

- Now *both* intercept and slope vary by group
- Effect of x differs across schools / countries / firms
- Heterogeneity is explicitly modeled, not averaged away
- Useful when you expect the effect itself to depend on context

Fitting multilevel models in R

- Frequentist / REML: `lme4`
  - → `lmer(y ~ x + (1 | group), data = d)` — random intercept
  - → `lmer(y ~ x + (1 + x | group), data = d)` — random slope
- Bayesian: `brms` or `rstanarm`
  - → Same formula syntax, full posterior draws
- Visualization: `sjPlot::plot_model()`, `broom.mixed`

Multilevel vs. fixed effects

	Multilevel	**Fixed effects**
Goal	Describe variation across groups	Remove group-level confounding
Assumes	Group effects uncorrelated with x	Nothing about that
Small groups	Handles well (shrinkage)	Noisy
Causal setting	Use if assumption plausible	Safer default

- If your goal is descriptive and you have many small groups, multilevel is natural
- If your goal is causal and you worry about group-level confounding, FE is safer

Multilevel: where to go next

- Gelman & Hill (2007) *Data Analysis Using Regression and Multilevel/Hierarchical Models*
- Gelman, Hill & Vehtari (2021) *Regression and Other Stories*, ch. 21–22
- McElreath *Statistical Rethinking*, ch. 13–14
- Bolker's `GLMM FAQ`: bbolker.github.io/mixedmodels-misc/glmmFAQ.html

Roadmap

Appendix: Multilevel models

Appendix: Text as data

Why text?

- Political science has always worked with text

Why text?

- Political science has always worked with text
  - → Speeches, manifestos, news, laws, tweets, transcripts

Why text?

- Political science has always worked with text
  - → Speeches, manifestos, news, laws, tweets, transcripts
- Historically: close reading and hand-coding

Why text?

- Political science has always worked with text
  - → Speeches, manifestos, news, laws, tweets, transcripts
- Historically: close reading and hand-coding
- Now: the volume makes automation essential

Why text?

- Political science has always worked with text
  - → Speeches, manifestos, news, laws, tweets, transcripts
- Historically: close reading and hand-coding
- Now: the volume makes automation essential
- A quantitative approach: turn documents into features, treat them like any other variable

Preprocessing

- Tokenization: split text into words or sub-words

Preprocessing

- Tokenization: split text into words or sub-words
- Lowercasing

Preprocessing

- Tokenization: split text into words or sub-words
- Lowercasing
- Remove punctuation, numbers, stopwords ("the", "and")

Preprocessing

- Tokenization: split text into words or sub-words
- Lowercasing
- Remove punctuation, numbers, stopwords (“the”, ”and”)
- Stemming / lemmatization: “running”, ”ran”, “runs” $\rightarrow$ ”run”

Preprocessing

- Tokenization: split text into words or sub-words
- Lowercasing
- Remove punctuation, numbers, stopwords (“the”, ”and”)
- Stemming / lemmatization: “running”, ”ran”, “runs” $\rightarrow$ ”run”
- Every decision changes the downstream analysis

Preprocessing

- Tokenization: split text into words or sub-words
- Lowercasing
- Remove punctuation, numbers, stopwords (“the”, ”and”)
- Stemming / lemmatization: “running”, ”ran”, “runs” $\rightarrow$ ”run”
- Every decision changes the downstream analysis
- Document these choices the way you would any other transformation

Bag of words

- Simplest representation: count how often each word appears

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**
  - → Rows: documents

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**
  - → Rows: documents
  - → Columns: unique words

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**
  - → Rows: documents
  - → Columns: unique words
  - → Cells: counts (or TF-IDF weights)

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**
  - → Rows: documents
  - → Columns: unique words
  - → Cells: counts (or TF-IDF weights)
- Ignores word order and grammar

Bag of words

- Simplest representation: count how often each word appears
- Build a **document-term matrix (DTM)**
  - → Rows: documents
  - → Columns: unique words
  - → Cells: counts (or TF-IDF weights)
- Ignores word order and grammar
- Surprisingly effective for many tasks

Dictionary methods

- Define a list of words associated with a concept

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...
  - → Populism: “elite”, ”people”, “corrupt”...

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...
  - → Populism: “elite”, ”people”, “corrupt”...
- Count occurrences per document, maybe normalize

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...
  - → Populism: “elite”, ”people”, “corrupt”...
- Count occurrences per document, maybe normalize
- Pro: simple, transparent, reproducible

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...
  - → Populism: “elite”, ”people”, “corrupt”...
- Count occurrences per document, maybe normalize
- Pro: simple, transparent, reproducible
- Con: brittle to context, sarcasm, domain shift

Dictionary methods

- Define a list of words associated with a concept
  - → Sentiment: “good”, ”great”, “bad”...
  - → Populism: “elite”, ”people”, “corrupt”...
- Count occurrences per document, maybe normalize
- Pro: simple, transparent, reproducible
- Con: brittle to context, sarcasm, domain shift
- Classic in political science — see Laver, Benoit & Garry (2003) Wordscores; LIWC

Topic models

- Unsupervised: discover themes in a corpus

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants
  - $\rightarrow$ Each document is a mixture of topics

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants
  - → Each document is a mixture of topics
  - → Each topic is a distribution over words

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants
  - → Each document is a mixture of topics
  - → Each topic is a distribution over words
- Structural Topic Model (STM, Roberts et al.)

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants
  - → Each document is a mixture of topics
  - → Each topic is a distribution over words
- Structural Topic Model (STM, Roberts et al.)
  - → Includes document-level covariates (author, year, party)

Topic models

- Unsupervised: discover themes in a corpus
- Latent Dirichlet Allocation (LDA) and descendants
  - → Each document is a mixture of topics
  - → Each topic is a distribution over words
- Structural Topic Model (STM, Roberts et al.)
  - → Includes document-level covariates (author, year, party)
- R packages: `topicmodels`, `stm`

Supervised classification

- You have labels for some documents; predict labels for the rest

Supervised classification

- You have labels for some documents; predict labels for the rest
- Examples:

Supervised classification

- You have labels for some documents; predict labels for the rest
- Examples:
  - → Is this speech about foreign policy? (yes / no)

Supervised classification

- You have labels for some documents; predict labels for the rest
- Examples:
  - → Is this speech about foreign policy? (yes / no)
  - → Is this tweet hate speech? (multi-class)

Supervised classification

- You have labels for some documents; predict labels for the rest
- Examples:
  - → Is this speech about foreign policy? (yes / no)
  - → Is this tweet hate speech? (multi-class)
- Pipeline: DTM or embeddings → classifier (logistic, SVM, random forest)

Supervised classification

- You have labels for some documents; predict labels for the rest
- Examples:
  - → Is this speech about foreign policy? (yes / no)
  - → Is this tweet hate speech? (multi-class)
- Pipeline: DTM or embeddings → classifier (logistic, SVM, random forest)
- Evaluation: held-out set, cross-validation, F1/precision/recall

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...
- Captures word *similarity* ("king" - "man" + "woman" $\approx$ "queen")

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...
- Captures word *similarity* ("king" - "man" + "woman" $\approx$ "queen")
- Large language models (GPT, Llama, etc.)

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...
- Captures word *similarity* ("king" - "man" + "woman" $\approx$ "queen")
- Large language models (GPT, Llama, etc.)
  - $\rightarrow$ Zero-shot classification, extraction, annotation

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...
- Captures word *similarity* ("king" - "man" + "woman" $\approx$ "queen")
- Large language models (GPT, Llama, etc.)
  - $\rightarrow$ Zero-shot classification, extraction, annotation
  - $\rightarrow$ Replacing hand-coders for many tasks

Embeddings and LLMs

- Represent words (or documents) as dense vectors that encode meaning
- word2vec, GloVe, sentence-BERT...
- Captures word *similarity* ("king" - "man" + "woman" $\approx$ "queen")
- Large language models (GPT, Llama, etc.)
  - $\rightarrow$ Zero-shot classification, extraction, annotation
  - $\rightarrow$ Replacing hand-coders for many tasks
  - $\rightarrow$ But: black box, bias, replication concerns

Text as data: where to go next

- Grimmer, Roberts & Stewart (2022) *Text as Data*. Princeton UP
  - → The standard reference for social scientists
- R packages worth knowing:
  - → `quanteda` — full preprocessing & analysis pipeline
  - → `stm` — structural topic models
  - → `text` — interface to transformer embeddings
- Course: Pablo Barberá's text-as-data materials (online)