5. Challenges and Tools in Empirical Work

PhD Applied Methods

Duncan Webb
NovaSBE

Spring 2026

Empirical work in practice

- In the previous lectures, we focused on the **identification** of causal effects
  - Potential outcomes framework
  - Randomized controlled trials
  - Selection bias and how to avoid it

Empirical work in practice

- In the previous lectures, we focused on the **identification** of causal effects
  - Potential outcomes framework
  - Randomized controlled trials
  - Selection bias and how to avoid it
- But implementing empirical research involves many additional **practical challenges**
- Today: focus on a set of tools and techniques for dealing with these challenges

Why focus on RCTs?

- We'll use **randomized controlled trials** as our main example throughout
- **Why?** Because with RCTs we can abstract away from identification concerns

Why focus on RCTs?

- We'll use **randomized controlled trials** as our main example throughout
- **Why?** Because with RCTs we can abstract away from identification concerns
- This allows us to focus on the **empirical difficulties** that arise even when identification is clean
- **Important**: Most of these tools apply to other research designs too (DiD, IV, RDD, etc.)

Overview of topics

1. **Choosing controls**: Which variables should we control for?
2. **Heterogeneous treatment effects**: How do effects vary across groups?
3. **Attrition**: What to do when participants drop out?
4. **Multiple hypothesis testing**: How to avoid false discoveries?
5. **Robustness and replication**: How reliable are empirical results?
6. **Standard errors**: How to correctly compute uncertainty?
7. **Spillovers**: What if SUTVA is violated?
8. **Measurement**: How to deal with measurement error?

Why add control variables?

Two main reasons to add control variables in a regression:

1. For identification of the causal effect

- Control for confounders
- Make the conditional independence assumption more plausible

Why add control variables?

Two main reasons to add control variables in a regression:

1. For identification of the causal effect

- Control for confounders
- Make the conditional independence assumption more plausible

2. To reduce variance of the error term

- Increase precision of estimates
- Reduce standard errors
- Increase statistical power

Why add control variables?

Two main reasons to add control variables in a regression:

1. For identification of the causal effect

- Control for confounders
- Make the conditional independence assumption more plausible

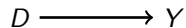
2. To reduce variance of the error term

- Increase precision of estimates
- Reduce standard errors
- Increase statistical power

With a large set of potential controls, **which should we choose?**

Identification through DAGs

- **Directed Acyclic Graphs** (DAGs): graphical representation of causal relationships
- We can encode the relationship between D and Y using an arrow
- Direction emphasizes that D causes Y, not vice versa
- More intuitive than equations for complex settings

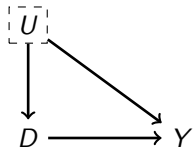


Confounders

- U affects both treatment D and outcome Y
- U is a **confounder**
- Creates association between D and Y that is *not* the causal effect
- Why? D and Y correlated through their common cause U

Confounders

- U affects both treatment D and outcome Y
- U is a **confounder**
- Creates association between D and Y that is *not* the causal effect
- Why? D and Y correlated through their common cause U



Example: Education (D) and wages (Y)
Ability (U) affects both choices
High ability $\implies$ more education *and* higher wages
$\implies$ Observe D and Y correlated even without causal effect

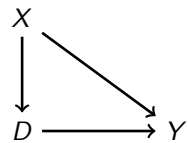
Controlling for confounders

- If X is observable confounder, can control for it
- Compare $D = 1$ vs. $D = 0$ *within* values of X
- Removes spurious association through X
- Recovers causal effect of D on Y

Controlling for confounders

- If X is observable confounder, can control for it
- Compare $D = 1$ vs. $D = 0$ *within* values of X
- Removes spurious association through X
- Recovers causal effect of D on Y

Example continued:
Control for ability (or test scores as proxy)
Compare high-ability people with/without education
Compare low-ability people with/without education
$\Longrightarrow$ Removes confounding by ability



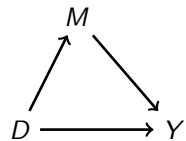
Conditioning on X makes D and Y independent (conditional on X)

Bad controls: Post-treatment variables

- M is caused by treatment D
- Also called “mediator”
- Controlling for M blocks part of causal effect
- Shuts down mechanism: $D \rightarrow M \rightarrow Y$

Bad controls: Post-treatment variables

- M is caused by treatment D
- Also called “mediator”
- Controlling for M blocks part of causal effect
- Shuts down mechanism: $D \to M \to Y$



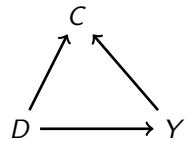
Example: Job training $\to$ skills $\to$ earnings
Controlling for skills misses indirect effect

Bad controls: Colliders

- C is caused by *both* D and Y
- Called a “collider”
- D and Y are independent
- But if you condition on C, creates spurious correlation

Bad controls: Colliders

- C is caused by *both* D and Y
- Called a "collider"
- D and Y are independent
- But if you condition on C, creates spurious correlation



Intuition:
Knowing C gives information about D and Y
If D high but C low, can infer Y must be low
$\implies$ Conditioning on C makes D and Y correlated
Even though no causal relationship!

Collider example: Beauty and talent

Setting: Study beauty (D) and talent (Y) among actors

- Beauty increases probability of becoming actor
- Talent increases probability of becoming actor
- Sample: only people who became actors

Collider example: Beauty and talent

Setting: Study beauty (D) and talent (Y) among actors

- Beauty increases probability of becoming actor
- Talent increases probability of becoming actor
- Sample: only people who became actors

Problem: "Becoming an actor" is a collider

- Both beauty and talent cause it
- By conditioning on actors, we create spurious correlation

Collider example: Beauty and talent

Setting: Study beauty (D) and talent (Y) among actors

- Beauty increases probability of becoming actor
- Talent increases probability of becoming actor
- Sample: only people who became actors

Problem: "Becoming an actor" is a collider

- Both beauty and talent cause it
- By conditioning on actors, we create spurious correlation

Result: Among actors, beauty and talent negatively correlated!

- Very beautiful actors succeed with low talent
- Less beautiful actors need high talent to succeed
- This correlation is *not* causal - it's an artifact of sample selection

Summary: Which controls?

Type	Control?	Why?
Confounder	Yes	Removes confounding
Post-treatment	No	Blocks causal mechanism
Collider	No	Creates spurious correlation

Key takeaway: Not all controls are good! Need to think carefully about the causal structure

Controls for precision

Even in RCTs (where identification is clean), controls increase precision

Recall: $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$ where σ^2 is residual variance

If X_i predicts Y_i, then including controls reduces σ^2

$\implies$ Lower standard errors, higher power

TABLE V

OLS and Reduced-Form Estimates of Effect of Class-Size Assignment on Average Percentile of Stanford Achievement Test

Explanatory variable	Reduced form: initial class size (5)	(6)	(7)	(8)
Small class	4.82	5.37	5.36	5.37
	(2.19)	(1.25)	(1.21)	(1.19)
Regular/aide class	.12	.29	.53	.31
	(2.23)	(1.13)	(1.09)	(1.07)
White/Asian (1 = yes	—	—	8.35	8.44
			(1.35)	(1.36)
Girl (1 = yes)	—	—	4.48	4.39
			(.63)	(.63)
Free lunch (1 = yes)	—	—	−13.15	−13.07
			(.77)	(.77)
White teacher	—	—	—	−.57
				(2.10)
Teacher experience	—	—	—	.26
				(.10)
Master's degree	—	—	—	−.51
				(1.06)
School fixed effects	No	Yes	Yes	Yes
R^2	.01	.25	.31	.31

STAR experiment: Controls reduce SEs, point estimates similar

Which controls for precision?

Ideal controls:

- Strongly correlated with outcome Y_i
- Pre-treatment (measured before randomization)
- Not affected by treatment

Which controls for precision?

Ideal controls:

- Strongly correlated with outcome Y_i
- Pre-treatment (measured before randomization)
- Not affected by treatment

Common choices:

- Baseline outcome (e.g., pre-treatment earnings)
- Demographics (age, gender, education)
- Stratification variables

Which controls for precision?

Ideal controls:

- Strongly correlated with outcome Y_i
- Pre-treatment (measured before randomization)
- Not affected by treatment

Common choices:

- Baseline outcome (e.g., pre-treatment earnings)
- Demographics (age, gender, education)
- Stratification variables

With many potential controls, how to choose?

Control selection with LASSO (for RCTs)

Post-Double-Selection LASSO (Belloni et al., 2014)

Procedure:

1. Run LASSO of Y_i on all controls (select predictors of outcome)
2. Run LASSO of D_i on all controls (select predictors of treatment)
3. Final regression: union of variables from steps 1 and 2

Control selection with LASSO (for RCTs)

Post-Double-Selection LASSO (Belloni et al., 2014)

Procedure:

1. Run LASSO of Y_i on all controls (select predictors of outcome)
2. Run LASSO of D_i on all controls (select predictors of treatment)
3. Final regression: union of variables from steps 1 and 2

LASSO: penalized regression that automatically sets some coefficients to zero

$$\min_{\beta,\gamma} \sum_{i=1}^{n} (Y_i - \beta D_i - X_i'\gamma)^2 + \lambda \sum_{j=1}^{p} |\gamma_j| \tag{1}$$

Control selection with LASSO (for RCTs)

Post-Double-Selection LASSO (Belloni et al., 2014)

Procedure:

1. Run LASSO of Y_i on all controls (select predictors of outcome)
2. Run LASSO of D_i on all controls (select predictors of treatment)
3. Final regression: union of variables from steps 1 and 2

LASSO: penalized regression that automatically sets some coefficients to zero

$$\min_{\beta,\gamma} \sum_{i=1}^{n} (Y_i - \beta D_i - X_i'\gamma)^2 + \lambda \sum_{j=1}^{p} |\gamma_j| \tag{1}$$

Advantages: Data-driven, reduces researcher discretion, helps avoid p-hacking

Outline

Beyond average treatment effects

So far: focused on **average treatment effect** (ATE)

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

Beyond average treatment effects

So far: focused on **average treatment effect** (ATE)

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

But effects likely vary across individuals/groups

Conditional Average Treatment Effect (CATE):

$$\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i] \tag{3}$$

where X_i are observable characteristics

Beyond average treatment effects

So far: focused on **average treatment effect** (ATE)

$$\tau_{ATE} = \mathbb{E}[Y_i(1) - Y_i(0)] \tag{2}$$

But effects likely vary across individuals/groups

Conditional Average Treatment Effect (CATE):

$$\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i] \tag{3}$$

where X_i are observable characteristics

Question: Why care about heterogeneity?

Why study heterogeneous effects?

1. **Policy targeting**
   - Which subgroups benefit most from the treatment?
   - Can improve cost-effectiveness

Why study heterogeneous effects?

1. Policy targeting

- Which subgroups benefit most from the treatment?
- Can improve cost-effectiveness

2. Understanding mechanisms

- How does the treatment work?
- Who is most affected and why?

Why study heterogeneous effects?

1. Policy targeting

- Which subgroups benefit most from the treatment?
- Can improve cost-effectiveness

2. Understanding mechanisms

- How does the treatment work?
- Who is most affected and why?

3. External validity

- Will treatment work in different contexts?
- What characteristics predict larger effects?

Basic approach: Interaction terms

Simple regression with interactions:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \delta(D_i \times X_i) + u_i \tag{4}$$

where X_i is characteristic of interest

Basic approach: Interaction terms

Simple regression with interactions:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \delta(D_i \times X_i) + u_i \tag{4}$$

where X_i is characteristic of interest

Interpretation:

- β: effect for group with $X_i = 0$
- $\beta + \delta$: effect for group with $X_i = 1$
- δ: *differential* effect across groups

Basic approach: Interaction terms

Simple regression with interactions:

$$Y_i = \alpha + \beta D_i + \gamma X_i + \delta(D_i \times X_i) + u_i \tag{4}$$

where X_i is characteristic of interest

Interpretation:

- β: effect for group with $X_i = 0$
- $\beta + \delta$: effect for group with $X_i = 1$
- δ: *differential* effect across groups

Can test: $H_0 : \delta = 0$ (no heterogeneity)

Example: Cash transfers and suicide (Roth et al. 2019)

Setting: Rollout of unconditional cash transfer in Indonesia
Treatment: ˜10% of household expenditure
Outcome: Suicide rate per 100,000 people

Example: Cash transfers and suicide (Roth et al. 2019)

Setting: Rollout of unconditional cash transfer in Indonesia
Treatment: ˜10% of household expenditure
Outcome: Suicide rate per 100,000 people

Main result: Transfer reduced suicides by 0.36 per 100,000 (18% reduction)

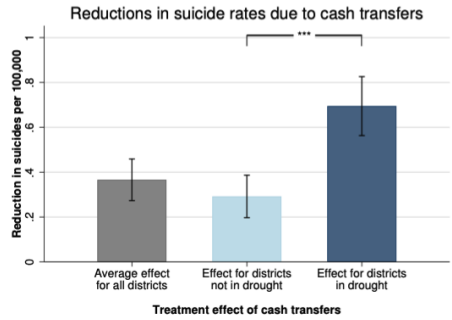
Example: Cash transfers and suicide (Roth et al. 2019)

Setting: Rollout of unconditional cash transfer in Indonesia
Treatment: ˜10% of household expenditure
Outcome: Suicide rate per 100,000 people

Main result: Transfer reduced suicides by 0.36 per 100,000 (18% reduction)

Heterogeneity:

- Larger reductions in areas with economic distress
- Larger effects in areas experiencing drought
- Benefits greatest among most vulnerable



Challenges with interaction terms

Problem 1: Many potential dimensions of heterogeneity

- Age, gender, income, education, location, etc.
- Which interactions to test?

Challenges with interaction terms

Problem 1: Many potential dimensions of heterogeneity

- Age, gender, income, education, location, etc.
- Which interactions to test?

Problem 2: Multiple hypothesis testing

- Test many interactions $\implies$ some significant by chance
- Risk of p-hacking / false discoveries

Challenges with interaction terms

Problem 1: Many potential dimensions of heterogeneity

- Age, gender, income, education, location, etc.
- Which interactions to test?

Problem 2: Multiple hypothesis testing

- Test many interactions $\implies$ some significant by chance
- Risk of p-hacking / false discoveries

$\implies$ Need more disciplined, data-driven approach

Machine learning for CATEs

Goal: Estimate $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ in data-driven way

Two fundamental challenges:

Machine learning for CATEs

Goal: Estimate $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ in data-driven way

Two fundamental challenges:

1. The "ground truth" problem

- Standard ML: Cross-validate by comparing $\hat{Y}_i$ to observed Y_i
- Causal inference: We *never* observe $\tau_i = Y_i(1) - Y_i(0)$ for any unit
- $\implies$ Cannot directly cross-validate treatment effect predictions

Machine learning for CATEs

Goal: Estimate $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$ in data-driven way

Two fundamental challenges:

1. The "ground truth" problem

- Standard ML: Cross-validate by comparing $\hat{Y}_i$ to observed Y_i
- Causal inference: We *never* observe $\tau_i = Y_i(1) - Y_i(0)$ for any unit
- $\implies$ Cannot directly cross-validate treatment effect predictions

2. The "adaptive bias" problem

- Standard CART uses same data to: (1) find splits, (2) estimate effects
- You split *because* $\hat{\tau}$ looked large $\implies$ estimates biased upward
- Like picking the best-performing stock, then reporting its past returns as expected future returns

Causal trees (Athey & Imbens 2016)

Idea: Partition covariate space to maximize treatment effect heterogeneity

"Honest" estimation (key innovation):

1. Split sample in half: training + estimation
2. Use training data to build tree (decide where to split)
3. Use estimation data to compute treatment effects within each leaf

Causal trees (Athey & Imbens 2016)

Idea: Partition covariate space to maximize treatment effect heterogeneity

"Honest" estimation (key innovation):

1. Split sample in half: training + estimation
2. Use training data to build tree (decide where to split)
3. Use estimation data to compute treatment effects within each leaf

Why this works: Estimation sample wasn't used to choose splits

$\Longrightarrow$ No cherry-picking bias, valid confidence intervals

Causal trees (Athey & Imbens 2016)

Idea: Partition covariate space to maximize treatment effect heterogeneity

"Honest" estimation (key innovation):

1. Split sample in half: training + estimation
2. Use training data to build tree (decide where to split)
3. Use estimation data to compute treatment effects within each leaf

Why this works: Estimation sample wasn't used to choose splits

$\Longrightarrow$ No cherry-picking bias, valid confidence intervals

Splitting criterion: Find partitions where treatment effects differ most across leaves, but penalize tiny subgroups (noisy estimates)

Causal trees (Athey & Imbens 2016)

Idea: Partition covariate space to maximize treatment effect heterogeneity

"Honest" estimation (key innovation):

1. Split sample in half: training + estimation
2. Use training data to build tree (decide where to split)
3. Use estimation data to compute treatment effects within each leaf

Why this works: Estimation sample wasn't used to choose splits

$\Longrightarrow$ No cherry-picking bias, valid confidence intervals

Splitting criterion: Find partitions where treatment effects differ most across leaves, but penalize tiny subgroups (noisy estimates)

Causal forests (Wager & Athey 2018): Average many honest trees for smoother estimates

GATES: Group Average Treatment Effects

The problem: Many covariates, don't know *ex ante* which matter for heterogeneity
Key insight (Chernozhukov et al. 2020): Don't estimate the full CATE—estimate **features** of it

GATES: Group Average Treatment Effects

The problem: Many covariates, don't know *ex ante* which matter for heterogeneity
Key insight (Chernozhukov et al. 2020): Don't estimate the full CATE—estimate **features** of it
GATES procedure:

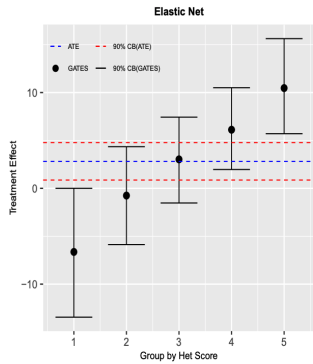
1. Use *any* ML to predict $\hat{\tau}(X_i)$
2. Sort by predicted effect
3. Group into quintiles
4. Estimate actual ATE within each group

GATES: Group Average Treatment Effects

The problem: Many covariates, don't know *ex ante* which matter for heterogeneity
Key insight (Chernozhukov et al. 2020): Don't estimate the full CATE—estimate **features** of it
GATES procedure:

1. Use *any* ML to predict $\hat{\tau}(X_i)$
2. Sort by predicted effect
3. Group into quintiles
4. Estimate actual ATE within each group

Interpretation: If ML detects real heterogeneity, Group 5 should have larger effects than Group 1



Upward slope $\Longrightarrow$ ML captures real heterogeneity

CLAN: Who are the high and low responders?

CLAN (Classification Analysis):
Once GATES reveals heterogeneity, **who** are high vs. low responders?

Method: Compare characteristics of "most affected" (G_K) vs. "least affected" (G_1) groups

CLAN: Who are the high and low responders?

CLAN (Classification Analysis):
Once GATES reveals heterogeneity, **who** are high vs. low responders?

Method: Compare characteristics of "most affected" (G_K) vs. "least affected" (G_1) groups

Example: Immunization in Haryana, India (Banerjee et al. 2019)

- RCT of SMS + incentives to boost vaccination
- Policy question: *where* to target expensive intervention?

CLAN: Who are the high and low responders?

CLAN (Classification Analysis):
Once GATES reveals heterogeneity, **who** are high vs. low responders?

Method: Compare characteristics of "most affected" (G_K) vs. "least affected" (G_1) groups

Example: Immunization in Haryana, India (Banerjee et al. 2019)

- RCT of SMS + incentives to boost vaccination
- Policy question: *where* to target expensive intervention?

CLAN finding: Villages with low baseline rates benefited most $\implies$ target rollout there

TABLE 5. CLAN of Immunization Incentives

	Elastic Net		
	20% Most (δ_5)	20% Least (δ_1)	Difference ($\delta_5 - \delta_1$)
Number of vaccines to pregnant mother	2.161 (2.110,2.212) -	2.288 (2.237,2.337) -	-0.128 (-0.200,-0.055) [0.001]
Number of vaccines to child since birth	4.230 (4.100,4.369) -	4.714 (4.573,4.860) -	-0.513 (-0.710,-0.311) [0.000]
Fraction of children received polio drops	1.000 (1.000,1.000) -	1.000 (1.000,1.000) -	0.000 (0.000,0.000) [0.000]
Number of polio drops to child	2.964 (2.954,2.975) -	2.998 (2.987,3.007) -	-0.033 (-0.047,-0.019) [0.000]
Fraction of children received immunization card	0.899 (0.878,0.922) -	0.932 (0.908,0.956) -	-0.036 (-0.065,-0.004) [0.000]
Fraction of children received Measles vaccine by 15 months of age	0.127 (0.100,0.155) -	0.255 (0.230,0.282) -	-0.131 (-0.167,-0.094) [0.052]
Fraction of children received Measles at credible locations	0.290 (0.252,0.327) -	0.435 (0.400,0.470) -	-0.152 (-0.198,-0.097) [0.000]

Notes: Medians over 100 splits. 90% confidence interval in parenthesis.
Notes: P-values for the hypothesis that the parameter is equal to zero in brackets.

Top 20% vs. bottom 20% responders

Outline

The problem of attrition

Attrition: Participants drop out between treatment assignment and outcome measurement

Common in:

- Surveys with multiple waves
- RCTs
- Any study with follow-up measurement

The problem of attrition

Attrition: Participants drop out between treatment assignment and outcome measurement

Common in:

- Surveys with multiple waves
- RCTs
- Any study with follow-up measurement

Why is this a problem?

Even with perfect randomization, attrition can create selection bias

Example: Job training RCT

Setup: Randomize 1,000 people to training vs. control

Measure employment 1 year later

Example: Job training RCT

Setup: Randomize 1,000 people to training vs. control

Measure employment 1 year later

Problem: 20% of treatment group, 10% of control don't respond

Example: Job training RCT

Setup: Randomize 1,000 people to training vs. control

Measure employment 1 year later

Problem: 20% of treatment group, 10% of control don't respond

Issue: If attrition related to outcomes, estimates biased

- Maybe successful people in treatment don't respond (too busy working)
- Maybe unsuccessful people in control don't respond (discouraged)
- $\Longrightarrow$ Comparing non-random samples

Testing for differential attrition

First step: Test whether attrition differs by treatment status

Regress attrition indicator on treatment:

$$\text{Attrited}_i = \alpha + \beta D_i + u_i \tag{5}$$

Test $H_0 : \beta = 0$

Testing for differential attrition

First step: Test whether attrition differs by treatment status

Regress attrition indicator on treatment:

$$\text{Attrited}_i = \alpha + \beta D_i + u_i \tag{5}$$

Test $H_0 : \beta = 0$

If reject H_0: **differential attrition**

Treatment affects who stays in sample

Testing for differential attrition

First step: Test whether attrition differs by treatment status

Regress attrition indicator on treatment:

$$\text{Attrited}_i = \alpha + \beta D_i + u_i \tag{5}$$

Test $H_0 : \beta = 0$

If reject H_0: **differential attrition**

Treatment affects who stays in sample

But: Even if no differential attrition, can still have bias if attrition related to potential outcomes

Testing attrition on baseline characteristics

Additional test: Compare baseline characteristics of attriters vs. non-attriters

For each baseline covariate X_i:

$$X_i = \alpha + \beta \text{Attrited}_i + \gamma D_i + \delta(\text{Attrited}_i \times D_i) + u_i \quad (6)$$

Testing attrition on baseline characteristics

Additional test: Compare baseline characteristics of attriters vs. non-attriters

For each baseline covariate X_i:

$$X_i = \alpha + \beta \text{Attrited}_i + \gamma D_i + \delta(\text{Attrited}_i \times D_i) + u_i \tag{6}$$

Interpretation:

- $\beta \neq 0$: Attriters differ from non-attriters (potential bias)
- $\delta \neq 0$: **Type** of attritor differs by treatment $\implies$ confounds treatment estimates

Testing attrition on baseline characteristics

Additional test: Compare baseline characteristics of attriters vs. non-attriters

For each baseline covariate X_i:

$$X_i = \alpha + \beta \text{Attrited}_i + \gamma D_i + \delta(\text{Attrited}_i \times D_i) + u_i \tag{6}$$

Interpretation:

- $\beta \neq 0$: Attriters differ from non-attriters (potential bias)
- $\delta \neq 0$: **Type** of attritor differs by treatment $\implies$ confounds treatment estimates

Important: These are diagnostic tests, not solutions

Solutions to attrition

Three main approaches:

1. **Prevention** (best option)
   - Minimize attrition through study design
   - Multiple contact methods, incentives, tracking

Solutions to attrition

Three main approaches:

1. **Prevention** (best option)
   - Minimize attrition through study design
   - Multiple contact methods, incentives, tracking

2. **Bounding** (Lee bounds)
   - Worst-case scenario analysis
   - No parametric assumptions

Solutions to attrition

Three main approaches:

1. **Prevention** (best option)
   - Minimize attrition through study design
   - Multiple contact methods, incentives, tracking

2. **Bounding** (Lee bounds)
   - Worst-case scenario analysis
   - No parametric assumptions

3. **Modeling** (e.g., inverse probability weighting)
   - Reweight to correct for selection
   - Requires strong assumptions

Lee bounds: Intuition

Idea (Lee 2009): Bound treatment effect without knowing why people attrite

Key assumption: Monotonicity

- Treatment affects attrition in only one direction
- E.g., treatment only increases attrition (or only decreases it)

Lee bounds: Intuition

Idea (Lee 2009): Bound treatment effect without knowing why people attrite

Key assumption: Monotonicity

- Treatment affects attrition in only one direction
- E.g., treatment only increases attrition (or only decreases it)

Logic:

- Suppose treatment group has 80% response, control has 100%
- 20% of treatment group are "extra attriters" caused by treatment
- Worst case: these 20% had highest (or lowest) outcomes
- $\Longrightarrow$ Trim top/bottom 20% of treatment group to get bounds

Lee bounds: Method

Setup: Let $S_i(d) \in \{0, 1\}$ indicate whether observed if assigned treatment d

Monotonicity assumption: $S_i(1) \geq S_i(0)$ for all i (or vice versa)

Lee bounds: Method

Setup: Let $S_i(d) \in \{0, 1\}$ indicate whether observed if assigned treatment d

Monotonicity assumption: $S_i(1) \geq S_i(0)$ for all i (or vice versa)

Construction (Example: 80% response in treatment, 100% in control):

1. Identify excess proportion: 20% are treatment-induced attriters

Lee bounds: Method

Setup: Let $S_i(d) \in \{0, 1\}$ indicate whether observed if assigned treatment d

Monotonicity assumption: $S_i(1) \geq S_i(0)$ for all i (or vice versa)

Construction (Example: 80% response in treatment, 100% in control):

1. Identify excess proportion: 20% are treatment-induced attriters
2. Construct bounds by trimming:
   - **Upper bound**: Drop bottom 20% of treatment outcomes
   - **Lower bound**: Drop top 20% of treatment outcomes

Lee bounds: Method

Setup: Let $S_i(d) \in \{0, 1\}$ indicate whether observed if assigned treatment d

Monotonicity assumption: $S_i(1) \geq S_i(0)$ for all i (or vice versa)

Construction (Example: 80% response in treatment, 100% in control):

1. Identify excess proportion: 20% are treatment-induced attriters
2. Construct bounds by trimming:
   - **Upper bound**: Drop bottom 20% of treatment outcomes
   - **Lower bound**: Drop top 20% of treatment outcomes
3. Compare trimmed treatment mean to control mean

Lee bounds: Method

Setup: Let $S_i(d) \in \{0, 1\}$ indicate whether observed if assigned treatment d

Monotonicity assumption: $S_i(1) \geq S_i(0)$ for all i (or vice versa)

Construction (Example: 80% response in treatment, 100% in control):

1. Identify excess proportion: 20% are treatment-induced attriters
2. Construct bounds by trimming:
   - **Upper bound**: Drop bottom 20% of treatment outcomes
   - **Lower bound**: Drop top 20% of treatment outcomes
3. Compare trimmed treatment mean to control mean

Intuition: We don't know which 20% would have attrited, so consider worst cases

Manski bounds

Alternative approach (Manski 1990): Even weaker assumptions

For attriters, assume worst-case outcomes within support $[y_{\min}, y_{\max}]$:

Manski bounds

Alternative approach (Manski 1990): Even weaker assumptions

For attriters, assume worst-case outcomes within support $[y_{\min}, y_{\max}]$:

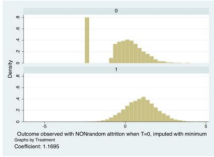
- **Upper bound**: Treatment attriters have $Y = y_{\max}$, control attriters have $Y = y_{\min}$
- **Lower bound**: Treatment attriters have $Y = y_{\min}$, control attriters have $Y = y_{\max}$

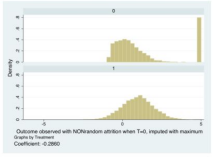
Manski bounds

Alternative approach (Manski 1990): Even weaker assumptions

For attriters, assume worst-case outcomes within support $[y_{\min}, y_{\max}]$:

- **Upper bound**: Treatment attriters have $Y = y_{\max}$, control attriters have $Y = y_{\min}$
- **Lower bound**: Treatment attriters have $Y = y_{\min}$, control attriters have $Y = y_{\max}$





Inverse probability weighting

Alternative approach: Model attrition probability

Basic intuition: Reweight observed sample to look like full sample (before attrition)

Inverse probability weighting

Alternative approach: Model attrition probability

Basic intuition: Reweight observed sample to look like full sample (before attrition)

Step 1: Estimate probability of non-attrition: $\hat{p}_i = \hat{\Pr}(S_i = 1 | X_i, D_i)$

using logit/probit with baseline covariates X_i

Inverse probability weighting

Alternative approach: Model attrition probability

Basic intuition: Reweight observed sample to look like full sample (before attrition)

Step 1: Estimate probability of non-attrition: $\hat{p}_i = \hat{\Pr}(S_i = 1 | X_i, D_i)$

using logit/probit with baseline covariates X_i

Step 2: Use sample weight $w_i = 1/\hat{p}_i$ in main regression

$$Y_i = \alpha + \tau D_i + X_i' \beta + (X_i \times D_i)' \gamma + \varepsilon_i \tag{7}$$

Inverse probability weighting

Alternative approach: Model attrition probability

Basic intuition: Reweight observed sample to look like full sample (before attrition)

Step 1: Estimate probability of non-attrition: $\hat{p}_i = \hat{\Pr}(S_i = 1 | X_i, D_i)$

using logit/probit with baseline covariates X_i

Step 2: Use sample weight $w_i = 1/\hat{p}_i$ in main regression

$$Y_i = \alpha + \tau D_i + X_i' \beta + (X_i \times D_i)' \gamma + \varepsilon_i \tag{7}$$

Assumption: Attrition independent of outcomes conditional on (X_i, D_i) — Much stronger than Lee bounds!

Outline

What is a p-value?

Question: What is the definition of a p-value?

What is a p-value?

Question: What is the definition of a p-value?

Answer: The probability of observing a result at least as extreme as what we got, *if there were no true effect*

Standard hypothesis testing assumes: **We only do one test**

What is a p-value?

Question: What is the definition of a p-value?

Answer: The probability of observing a result at least as extreme as what we got, *if there were no true effect*

Standard hypothesis testing assumes: **We only do one test**

Problem: What if we run multiple tests?

The multiple testing problem

Thought experiment: Run 1,000 regressions on completely random data with $Y_i \perp X_i$

With $\alpha = 0.05$, how many will look "significant"?

The multiple testing problem

Thought experiment: Run 1,000 regressions on completely random data with $Y_i \perp X_i$

With $\alpha = 0.05$, how many will look "significant"?

Answer: About 50 (i.e., 5% of 1,000)

The multiple testing problem

Thought experiment: Run 1,000 regressions on completely random data with $Y_i \perp X_i$

With $\alpha = 0.05$, how many will look "significant"?

Answer: About 50 (i.e., 5% of 1,000)

Even worse: What if we run 1,000 tests and *only report* the significant ones?

The multiple testing problem

Thought experiment: Run 1,000 regressions on completely random data with $Y_i \perp X_i$

With $\alpha = 0.05$, how many will look "significant"?

Answer: About 50 (i.e., 5% of 1,000)

Even worse: What if we run 1,000 tests and *only report* the significant ones?

$\Longrightarrow$ Our published results will be full of false positives!

Why this matters for empirical work

Multiple testing arises naturally in many contexts:

- Testing treatment effects on **many outcomes**
  - Health study: blood pressure, cholesterol, weight, ...
  - Education: test scores, attendance, graduation, ...

Why this matters for empirical work

Multiple testing arises naturally in many contexts:

- Testing treatment effects on **many outcomes**
  - Health study: blood pressure, cholesterol, weight, ...
  - Education: test scores, attendance, graduation, ...
- Testing for **heterogeneous effects**
  - By gender, age, income, region, ...

Why this matters for empirical work

Multiple testing arises naturally in many contexts:

- Testing treatment effects on **many outcomes**
  - Health study: blood pressure, cholesterol, weight, ...
  - Education: test scores, attendance, graduation, ...
- Testing for **heterogeneous effects**
  - By gender, age, income, region, ...
- Testing **many specifications**
  - Different control variables
  - Different samples
  - Different functional forms

Solutions: Pre-analysis plans

Pre-analysis plan (PAP): Document research design before seeing data

Specify in advance: primary outcomes, subgroup analyses, specifications

Solutions: Pre-analysis plans

Pre-analysis plan (PAP): Document research design before seeing data

Specify in advance: primary outcomes, subgroup analyses, specifications

Benefits: Prevents p-hacking, makes researcher choices transparent, now standard for RCTs

Solutions: Pre-analysis plans

Pre-analysis plan (PAP): Document research design before seeing data

Specify in advance: primary outcomes, subgroup analyses, specifications

Benefits: Prevents p-hacking, makes researcher choices transparent, now standard for RCTs

Limitation: Doesn't solve multiple testing itself—still need to adjust p-values

Solutions: Adjust p-values

Idea: Adjust significance thresholds to account for multiple testing

Bonferroni correction (classic method):

- Testing m hypotheses
- Reject if $p < \alpha/m$
- Example: 20 tests, want FWER $\leq 0.05 \implies$ use threshold 0.05/20 = 0.0025

Solutions: Adjust p-values

Idea: Adjust significance thresholds to account for multiple testing

Bonferroni correction (classic method):

- Testing m hypotheses
- Reject if $p < \alpha/m$
- Example: 20 tests, want FWER $\leq 0.05 \implies$ use threshold 0.05/20 = 0.0025

Problem: Very conservative

- Low power to detect true effects
- Especially with many tests

Family-Wise Error Rate (FWER)

Definition: Probability of making *at least one* false rejection

$$\text{FWER} = \Pr(\text{At least one false positive}) \tag{8}$$

Bonferroni controls FWER at level α

Family-Wise Error Rate (FWER)

Definition: Probability of making *at least one* false rejection

$$\text{FWER} = \Pr(\text{At least one false positive}) \tag{8}$$

Bonferroni controls FWER at level α

Issue: Very stringent criterion

- In exploratory research, may accept some false positives
- Want to balance false positives vs. false negatives

False Discovery Rate (FDR)

Definition: Expected proportion of false rejections among all rejections

$$\text{FDR} = \mathbb{E}\left[\frac{\text{False positives}}{\text{Total rejections}}\right] \tag{9}$$

Interpretation: Among all tests you reject, what fraction are false positives?

False Discovery Rate (FDR)

Definition: Expected proportion of false rejections among all rejections

$$\text{FDR} = \mathbb{E}\left[\frac{\text{False positives}}{\text{Total rejections}}\right] \tag{9}$$

Interpretation: Among all tests you reject, what fraction are false positives?

Difference from FWER:

- FWER: Probability of *any* false positive
- FDR: *Rate* of false positives among rejections
- FDR is less stringent $\implies$ more power

Benjamini-Hochberg procedure (1995)

Goal: Control FDR at level q (e.g., $q = 0.10$)

Procedure for m hypothesis tests:

1. **Sort p-values**: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$

Benjamini-Hochberg procedure (1995)

Goal: Control FDR at level q (e.g., $q = 0.10$)

Procedure for m hypothesis tests:

1. **Sort p-values**: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
2. **Find largest** k such that:

$$p_{(k)} \leq \frac{k}{m} \cdot q \tag{10}$$

Benjamini-Hochberg procedure (1995)

Goal: Control FDR at level q (e.g., $q = 0.10$)

Procedure for m hypothesis tests:

1. **Sort p-values**: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
2. **Find largest** k such that:

$$p_{(k)} \leq \frac{k}{m} \cdot q \tag{10}$$

3. **Reject all hypotheses** $H_{(1)}, \ldots, H_{(k)}$

Benjamini-Hochberg procedure (1995)

Goal: Control FDR at level q (e.g., $q = 0.10$)

Procedure for m hypothesis tests:

1. **Sort p-values**: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
2. **Find largest** k such that:

$$p_{(k)} \leq \frac{k}{m} \cdot q \tag{10}$$

3. **Reject all hypotheses** $H_{(1)}, \ldots, H_{(k)}$

Result: FDR $\leq q$ under independence

Benjamini-Hochberg: Intuition

Key insight: Look for p-values that are “too small” to be false positives

Under null: p-values uniformly distributed on [0,1]

Under alternative: p-values concentrated near 0

Benjamini-Hochberg: Intuition

Key insight: Look for p-values that are “too small” to be false positives

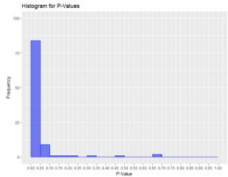
Under null: p-values uniformly distributed on [0,1]

Under alternative: p-values concentrated near 0

BH procedure finds cutoff where p-values deviate from uniform distribution

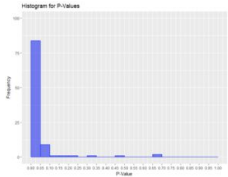
Benjamini-Hochberg: Visual intuition

True rejections: Concentrated near zero

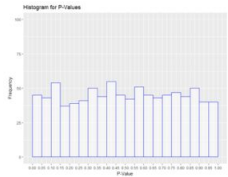


Benjamini-Hochberg: Visual intuition

True rejections: Concentrated near zero

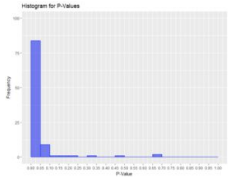


False rejections: Uniformly distributed

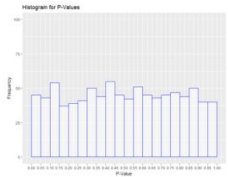


Benjamini-Hochberg: Visual intuition

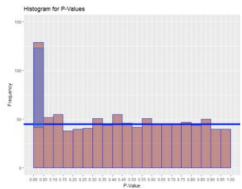
True rejections: Concentrated near zero



False rejections: Uniformly distributed



BH isolates p-values above uniform cutoff:



Example: Testing 10 hypotheses

Suppose we have 10 p-values, want FDR ≤ 0.10:

Rank k	$p_{(k)}$	$\frac{k}{m} \cdot q = \frac{k}{10} \cdot 0.10$	Reject?
1	0.001	0.010	✓
2	0.008	0.020	✓
3	0.015	0.030	✓
4	0.042	0.040	✓
5	0.056	0.050	×
6	0.120	0.060	×
⋮	⋮	⋮	×

Largest k with $p_{(k)} \leq k \cdot 0.01$ is $k = 4 \implies$ Reject first 4 hypotheses

Other approaches to reduce multiple testing

Beyond adjusting p-values:

1. Reduce number of tests / Use summary indices

- Pre-specify primary outcome(s)
- Use F-tests for joint hypotheses
- Combine outcomes into summary index (e.g., Anderson 2008)
  - Average standardized effects across related outcomes
  - Example: “Cognitive index” = average of test scores
  - Reduces m tests to 1 test

Other approaches to reduce multiple testing

Beyond adjusting p-values:

1. Reduce number of tests / Use summary indices

- Pre-specify primary outcome(s)
- Use F-tests for joint hypotheses
- Combine outcomes into summary index (e.g., Anderson 2008)
  - Average standardized effects across related outcomes
  - Example: "Cognitive index" = average of test scores
  - Reduces m tests to 1 test

2. Report all tests transparently

- Show results for all outcomes (not just significant ones)
- Let readers judge robustness

Practical recommendations

What should you do?

- **Pre-specify** main hypotheses when possible
- **Distinguish** confirmatory vs. exploratory analyses
- **Adjust p-values** when testing multiple outcomes
  - BH procedure for FDR control
  - Bonferroni for FWER control (if very conservative)
- **Report transparently**
  - Show all outcomes tested
  - Report both adjusted and unadjusted p-values
- **Implementation**: Available in R (`p.adjust()`) and Stata (`rwolf`, `wyoung`)

Outline

The replication crisis

Mounting evidence that empirical results are less robust than we thought

Problem 1: Low replication rates

- Psychology replication project: only 39% of studies replicate
- Economics: similar concerns emerging

The replication crisis

Mounting evidence that empirical results are less robust than we thought

Problem 1: Low replication rates

- Psychology replication project: only 39% of studies replicate
- Economics: similar concerns emerging

Problem 2: Researcher degrees of freedom

- Many choices in data analysis: sample, controls, specification, etc.
- Different choices $\implies$ different results
- Risk of p-hacking (intentional or not)

The replication crisis

Mounting evidence that empirical results are less robust than we thought

Problem 1: Low replication rates

- Psychology replication project: only 39% of studies replicate
- Economics: similar concerns emerging

Problem 2: Researcher degrees of freedom

- Many choices in data analysis: sample, controls, specification, etc.
- Different choices $\implies$ different results
- Risk of p-hacking (intentional or not)

Goal: Get to the truth! But what went wrong?

Evidence: Garden of forking paths

Study: Breznau, Rinke, Wuttke (PNAS 2022)

Design: 73 research teams test same hypothesis with same data

Does immigration reduce support for social policies?

Evidence: Garden of forking paths

Study: Breznau, Rinke, Wuttke (PNAS 2022)

Design: 73 research teams test same hypothesis with same data

Does immigration reduce support for social policies?

Results: Massive variance in estimates

- Estimates ranged from strongly negative to strongly positive
- Driven by: controls, fixed effects, clustering, sample choices
- Much variance unexplained even controlling for observables

Evidence: Researcher-driven variation

Study: Huntington-Klein et al.

Design: 146 economist teams estimate same effect with progressive restrictions

1. No restrictions (same data)
2. Specify DiD design
3. Pre-cleaned data

Evidence: Researcher-driven variation

Study: Huntington-Klein et al.

Design: 146 economist teams estimate same effect with progressive restrictions

1. No restrictions (same data)
2. Specify DiD design
3. Pre-cleaned data

Results:

- IQR of 3-4pp (avg effect: 4pp) with freedom
- IQR still 2pp with pre-cleaned data
- Both data and analytical choices matter

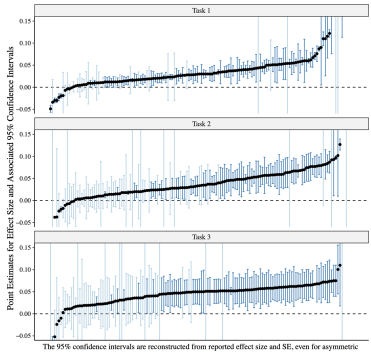


Figure 3: Specification Curve for All Reported Estimates by Task with Estimates Ordered From Smallest to Largest

What went wrong?

Three distinct but related issues:

1. Researcher degrees of freedom

- Many choices: sample definition, controls, specifications, outliers
- Different reasonable choices $\Longrightarrow$ different results
- Creates flexibility to find "significant" results (even unintentionally)

What went wrong?

Three distinct but related issues:

1. Researcher degrees of freedom

- Many choices: sample definition, controls, specifications, outliers
- Different reasonable choices $\Longrightarrow$ different results
- Creates flexibility to find "significant" results (even unintentionally)

2. Publication bias (editorial decisions)

- Journals prefer novel, significant, "clean" results
- Null results, replications rarely published (file drawer problem)
- $\Longrightarrow$ Published literature overrepresents significant findings

Solutions

How to improve reliability of empirical research?

1. Transparency

- Pre-register analysis plans (PAPs)
- Post data and code (replication packages)
- Report all specifications, not just significant ones

Solutions

How to improve reliability of empirical research?

1. Transparency

- Pre-register analysis plans (PAPs)
- Post data and code (replication packages)
- Report all specifications, not just significant ones

2. Robustness checks

- Show results not driven by one specific choice
- Vary sample, controls, functional forms
- Put extensive checks in appendix

Solutions

How to improve reliability of empirical research?

1. Transparency

- Pre-register analysis plans (PAPs)
- Post data and code (replication packages)
- Report all specifications, not just significant ones

2. Robustness checks

- Show results not driven by one specific choice
- Vary sample, controls, functional forms
- Put extensive checks in appendix

3. Design for power

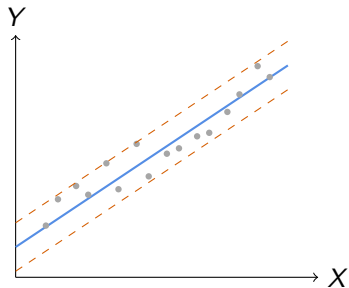
- Choose designs with high statistical power
- Reduces false negatives and winner's curse
- Better to detect true effects reliably

Outline

Homoskedasticity vs. Heteroskedasticity

Basic OLS assumption: Constant error variance (homoskedasticity)

Homoskedasticity

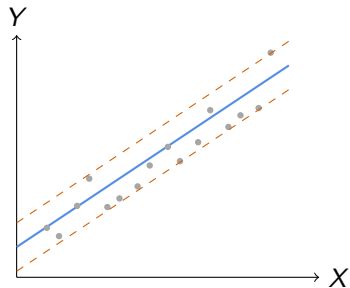


$V(\varepsilon_i|X_i) = \sigma^2$ (constant)

Homoskedasticity vs. Heteroskedasticity

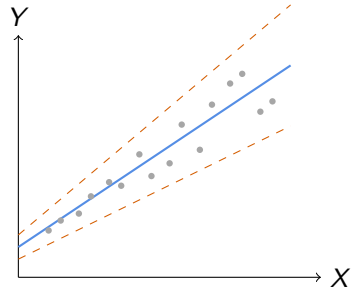
Basic OLS assumption: Constant error variance (homoskedasticity)

Homoskedasticity



$V(\varepsilon_i|X_i) = \sigma^2$ (constant)

Heteroskedasticity



$V(\varepsilon_i|X_i) = \sigma_i^2$ (varies)

Variance of OLS estimator

Under homoskedasticity:

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{11}$$

Variance of OLS estimator

Under homoskedasticity:

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{11}$$

Under heteroskedasticity:

$$V(\hat{\beta}) = (X'X)^{-1} X'\Omega X (X'X)^{-1} \tag{12}$$

where $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

Variance of OLS estimator

Under homoskedasticity:

$$V(\hat{\beta}) = \sigma^2 (X'X)^{-1} \tag{11}$$

Under heteroskedasticity:

$$V(\hat{\beta}) = (X'X)^{-1} X'\Omega X (X'X)^{-1} \tag{12}$$

where $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

Problem: If we use homoskedastic formula when heteroskedasticity present:

- Standard errors are wrong
- Confidence intervals and hypothesis tests invalid

Practical recommendation

Key point: There is almost never a good justification for assuming homoskedasticity

$\Longrightarrow$ **Always use robust (heteroskedasticity-consistent) standard errors**

Practical recommendation

Key point: There is almost never a good justification for assuming homoskedasticity

$\Longrightarrow$ **Always use robust (heteroskedasticity-consistent) standard errors**

Implementation:

- **Stata**: Add `, robust` option to regression
  - `reg y x, robust`

Practical recommendation

Key point: There is almost never a good justification for assuming homoskedasticity

$\Longrightarrow$ **Always use robust (heteroskedasticity-consistent) standard errors**

Implementation:

- **Stata**: Add `, robust` option to regression
  - `reg y x, robust`
- **R**: Use `vcovHC()` from `sandwich` package
  - `coeftest(model, vcov = vcovHC(model, type="HC1"))`

This should be your *default*, not an exception!

Clustering: Why does it matter?

Standard OLS assumption: Observations are independent (IID)

Problem: Often observations are correlated within groups

Clustering: Why does it matter?

Standard OLS assumption: Observations are independent (IID)

Problem: Often observations are correlated within groups

Common examples:

- **Panel data**: Same individual over time
- **Schools/classrooms**: Students in same school
- **Geographic/spatial**: Units in same region
- **Families**: Siblings, households

Clustering: Why does it matter?

Standard OLS assumption: Observations are independent (IID)

Problem: Often observations are correlated within groups

Common examples:

- **Panel data**: Same individual over time
- **Schools/classrooms**: Students in same school
- **Geographic/spatial**: Units in same region
- **Families**: Siblings, households

Key distinction: This is about *correlation in residuals*, not a SUTVA violation

Treatment of one unit doesn't affect others, but their errors are correlated

Clustered errors: Covariance structure

Example: 2 groups, 3 people each

Without clustering (IID)

$$\Omega = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

Diagonal: all errors independent

Clustered errors: Covariance structure

Example: 2 groups, 3 people each

Without clustering (IID)

$$\Omega = \begin{pmatrix} \sigma^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

Diagonal: all errors independent

With clustering

$$\Omega = \begin{pmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} & 0 & 0 & 0 \\ \sigma_{12} & \sigma^2 & \sigma_{23} & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma_{45} & \sigma_{46} \\ 0 & 0 & 0 & \sigma_{45} & \sigma^2 & \sigma_{56} \\ 0 & 0 & 0 & \sigma_{46} & \sigma_{56} & \sigma^2 \end{pmatrix}$$

Red: within-cluster correlations

Why clustering increases standard errors

Intuition: Correlation reduces effective sample size

Each observation provides *less independent information* when correlated

Why clustering increases standard errors

Intuition: Correlation reduces effective sample size

Each observation provides *less independent information* when correlated

Consequence: Variance of OLS estimator is actually higher:

$$V(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1} \tag{13}$$

where Ω has off-diagonal elements (covariances)

Why clustering increases standard errors

Intuition: Correlation reduces effective sample size

Each observation provides *less independent information* when correlated

Consequence: Variance of OLS estimator is actually higher:

$$V(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1} \tag{13}$$

where Ω has off-diagonal elements (covariances)

If you ignore clustering:

- Standard errors too small
- Over-rejection of null hypotheses
- False confidence in results

Modeling cluster structure: Random effects

Approach 1: Assume random group effects (panel data structure)

Error decomposes: $\varepsilon_{gi} = u_g + v_{gi}$ where u_g is group shock, v_{gi} is idiosyncratic

Modeling cluster structure: Random effects

Approach 1: Assume random group effects (panel data structure)

Error decomposes: $\varepsilon_{gi} = u_g + v_{gi}$ where u_g is group shock, v_{gi} is idiosyncratic

Then the covariance matrix Ω_g is:

$$\Omega_g = \begin{pmatrix} \sigma_u^2 + \sigma_v^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_u^2 \\ \sigma_u^2 & \cdots & \sigma_u^2 & \sigma_u^2 + \sigma_v^2 \end{pmatrix}$$

with $V(u_g) = \sigma_u^2$ and $V(v_{gi}) = \sigma_v^2$

Modeling cluster structure: Random effects

Approach 1: Assume random group effects (panel data structure)

Error decomposes: $\varepsilon_{gi} = u_g + v_{gi}$ where u_g is group shock, v_{gi} is idiosyncratic

Then the covariance matrix Ω_g is:

$$\Omega_g = \begin{pmatrix} \sigma_u^2 + \sigma_v^2 & \sigma_u^2 & \cdots & \sigma_u^2 \\ \sigma_u^2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_u^2 \\ \sigma_u^2 & \cdots & \sigma_u^2 & \sigma_u^2 + \sigma_v^2 \end{pmatrix}$$

with $V(u_g) = \sigma_u^2$ and $V(v_{gi}) = \sigma_v^2$

Note: $\forall i \neq j : \text{cov}(\varepsilon_{gi}, \varepsilon_{gj}) = \sigma_u^2$

Modeling cluster structure: Moulton factor

Approach 2: Constant within-group correlation (Moulton 1986)

Assumes within-group correlation is constant: $\text{cor}(\varepsilon_i, \varepsilon_j) = \rho$ if $G_i = G_j$, else 0

Modeling cluster structure: Moulton factor

Approach 2: Constant within-group correlation (Moulton 1986)

Assumes within-group correlation is constant: $\text{cor}(\varepsilon_i, \varepsilon_j) = \rho$ if $G_i = G_j$, else 0

Leading to covariance matrix: $\Omega_g = \sigma_e^2 \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{pmatrix}$

Modeling cluster structure: No assumptions

Approach 3: Don't assume specific structure—just estimate

$$\widehat{\Omega}_g = \begin{bmatrix} \hat{u}_{g1}^2 & \hat{u}_{g1}\hat{u}_{g2} & \hat{u}_{g1}\hat{u}_{g3} \\ \hat{u}_{g2}\hat{u}_{g1} & \hat{u}_{g2}^2 & \hat{u}_{g2}\hat{u}_{g3} \\ \hat{u}_{g3}\hat{u}_{g1} & \hat{u}_{g3}\hat{u}_{g2} & \hat{u}_{g3}^2 \end{bmatrix}$$

This is what **cluster-robust standard errors** do

Most flexible approach—recommended!

Practical recommendations: Clustering

When to cluster?

Whenever observations might be correlated within groups

How to choose cluster level?

- Cluster at level where treatment is assigned (if applicable)
- Cluster at highest level of potential correlation
- When in doubt: cluster at higher level (more conservative)

Practical recommendations: Clustering

When to cluster?

Whenever observations might be correlated within groups

How to choose cluster level?

- Cluster at level where treatment is assigned (if applicable)
- Cluster at highest level of potential correlation
- When in doubt: cluster at higher level (more conservative)

Implementation:

- **Stata**: `reg y x, cluster(groupvar)`
- **R**: `coeftest(model, vcov = vcovCL(model, cluster = ~groupvar))`

Result: Accounting for clustering *almost always* increases SEs

Outline

Spillovers as SUTVA violations

Recall SUTVA (Stable Unit Treatment Value Assumption):

Treatment of unit i doesn't affect outcomes of unit j

Spillovers as SUTVA violations

Recall SUTVA (Stable Unit Treatment Value Assumption):

Treatment of unit i doesn't affect outcomes of unit j

Spillovers: This assumption is violated!

Definition: Person i being treated affects outcome Y_j of person j

Spillovers as SUTVA violations

Recall SUTVA (Stable Unit Treatment Value Assumption):

Treatment of unit i doesn't affect outcomes of unit j

Spillovers: This assumption is violated!

Definition: Person i being treated affects outcome Y_j of person j

Examples:

- Vaccines: My vaccination protects you (positive spillover)
- Deworming: Treating some kids reduces transmission to others
- Information: Treated individuals share knowledge with control group
- Market effects: Training increases labor supply, affects wages

Indirect treatment effect

Object of interest: How much does treatment of others affect me?

Indirect Treatment Effect (ITE): Compare untreated in treated clusters vs. untreated in control clusters

$$\text{ITE} = \mathbb{E}[Y_i(0, D_{-i} = 1) - Y_i(0, D_{-i} = 0) | D_i = 0] \tag{14}$$

where:

- $Y_i(0, D_{-i} = 1)$: Outcome for untreated i when others in cluster are treated
- $Y_i(0, D_{-i} = 0)$: Outcome for untreated i when others in cluster are untreated

If ITE $\neq 0$, SUTVA is violated

Measuring spillovers: Two-step clustered RCT

Design:

1. **Step 1**: Randomize which clusters get treatment (e.g., schools, villages)

Measuring spillovers: Two-step clustered RCT

Design:

1. **Step 1**: Randomize which clusters get treatment (e.g., schools, villages)
2. **Step 2**: Randomize which individuals within treated clusters get treatment

Measuring spillovers: Two-step clustered RCT

Design:

1. **Step 1**: Randomize which clusters get treatment (e.g., schools, villages)
2. **Step 2**: Randomize which individuals within treated clusters get treatment

This allows estimation of:

- **Direct effect**: Treated vs. untreated in same cluster
- **Indirect effect**: Untreated in treated cluster vs. untreated in control cluster
- **Total effect**: Treated in treated cluster vs. untreated in control cluster

Measuring spillovers: Two-step clustered RCT

Design:

1. **Step 1**: Randomize which clusters get treatment (e.g., schools, villages)
2. **Step 2**: Randomize which individuals within treated clusters get treatment

This allows estimation of:

- **Direct effect**: Treated vs. untreated in same cluster
- **Indirect effect**: Untreated in treated cluster vs. untreated in control cluster
- **Total effect**: Treated in treated cluster vs. untreated in control cluster

Extension: Can vary proportion treated in each cluster (e.g., {0%, 30%, 60%})

Estimates how spillovers change with treatment intensity (but requires high power)

Example: Miguel & Kremer (2004) deworming study

Setting: Deworming treatment in Kenyan schools

Design: Phased randomization of schools

- Group 1 schools treated in 1998
- Group 2 schools treated in 1999
- Group 3 schools treated in 2000 (control)

Example: Miguel & Kremer (2004) deworming study

Setting: Deworming treatment in Kenyan schools

Design: Phased randomization of schools

- Group 1 schools treated in 1998
- Group 2 schools treated in 1999
- Group 3 schools treated in 2000 (control)

Key finding: Large positive spillovers

- Direct effect: Treated students healthier
- Spillover effect: Untreated students in nearby treated schools also healthier
- Mechanism: Reduced disease transmission

Example: Miguel & Kremer (2004) deworming study

Setting: Deworming treatment in Kenyan schools

Design: Phased randomization of schools

- Group 1 schools treated in 1998
- Group 2 schools treated in 1999
- Group 3 schools treated in 2000 (control)

Key finding: Large positive spillovers

- Direct effect: Treated students healthier
- Spillover effect: Untreated students in nearby treated schools also healthier
- Mechanism: Reduced disease transmission

Implication: Standard RCT would *underestimate* total benefits

(Control group also benefits from nearby treatment)

Testing for spillovers: The naive approach

Idea: Test if treatment effects vary by exposure to treated individuals

Basic regression:

$$Y_i = \alpha + \beta D_i + \gamma \text{Exposure}_i + \varepsilon_i \tag{15}$$

where Exposure_i = number of treated individuals near i

Problem with naive approach

Issue: Exposure is not exogenous!

Example problems:

- Urban vs. rural: People in dense cities more exposed to treated units
- Social networks: Popular people more exposed
- Geography: People near population centers more exposed

Problem with naive approach

Issue: Exposure is not exogenous!

Example problems:

- Urban vs. rural: People in dense cities more exposed to treated units
- Social networks: Popular people more exposed
- Geography: People near population centers more exposed

$\Longrightarrow$ Exposure$_i$ correlated with unobservables (e.g., density, centrality)

Problem with naive approach

Issue: Exposure is not exogenous!

Example problems:

- Urban vs. rural: People in dense cities more exposed to treated units
- Social networks: Popular people more exposed
- Geography: People near population centers more exposed

$\Longrightarrow$ Exposure$_i$ correlated with unobservables (e.g., density, centrality)

$\Longrightarrow$ γ is **biased** (OVB problem)

Cannot distinguish true spillovers from confounding by exposure-related factors

Solution: Recentering approach (Borusyak & Hull 2023)

Key idea: Control for *expected* exposure under random assignment

Steps:

1. Simulate random assignments many times (e.g., 10,000)
2. Calculate expected exposure: $\mathbb{E}[\text{Exposure}_i]$ for each unit

Solution: Recentering approach (Borusyak & Hull 2023)

Key idea: Control for *expected* exposure under random assignment

Steps:

1. Simulate random assignments many times (e.g., 10,000)
2. Calculate expected exposure: $\mathbb{E}[\text{Exposure}_i]$ for each unit
3. Regress outcome on actual exposure, controlling for expected exposure:

$$Y_i = \beta D_i + \gamma \text{Exposure}_i + \delta \mathbb{E}[\text{Exposure}_i] + \varepsilon_i \tag{16}$$

Recentering: Intuition

Why does this work?

- $\mathbb{E}[\text{Exposure}_i]$ captures confounders (density, centrality, etc.)

Recentering: Intuition

Why does this work?

- $\mathbb{E}[\text{Exposure}_i]$ captures confounders (density, centrality, etc.)
- These confounders affect *expected* exposure, not random deviations

Recentering: Intuition

Why does this work?

- $\mathbb{E}[\text{Exposure}_i]$ captures confounders (density, centrality, etc.)
- These confounders affect *expected* exposure, not random deviations
- Actual exposure $-$ expected exposure = exogenous variation from randomization

Recentering: Intuition

Why does this work?

- $\mathbb{E}[\text{Exposure}_i]$ captures confounders (density, centrality, etc.)
- These confounders affect *expected* exposure, not random deviations
- Actual exposure $-$ expected exposure = exogenous variation from randomization
- γ identifies causal spillover effect from this exogenous variation

Summary: Spillovers

Key takeaways:

- Spillovers violate SUTVA—treatment of i affects j
- Can be positive or negative, and quantitatively important
- **In RCTs**: Use two-step clustered design to measure indirect effects
- **Testing for spillovers**:
  - Naive exposure regressions are biased
  - Use recentering approach (Borusyak & Hull 2023)
  - Control for expected exposure to remove confounding
- Applies beyond RCTs: Any setting with spatial/network structure
- **Practical advice**: Always consider whether spillovers plausible in your setting

Outline

Measurement matters

Often overlooked: Measurement is critical but undervalued in econometrics courses

Key insight: Regardless of causal identification, if we don't measure the right thing, we can't even know Y_i!

Measurement matters

Often overlooked: Measurement is critical but undervalued in econometrics courses

Key insight: Regardless of causal identification, if we don't measure the right thing, we can't even know Y_i!

Two types of measurement error:

1. **Classical measurement error**: Random noise, independent of true value
   - $X_i^{\text{observed}} = X_i^{\text{true}} + \varepsilon_i$ where $\mathbb{E}[\varepsilon_i] = 0$, $\text{Cov}(X_i^{\text{true}}, \varepsilon_i) = 0$

Measurement matters

Often overlooked: Measurement is critical but undervalued in econometrics courses

Key insight: Regardless of causal identification, if we don't measure the right thing, we can't even know Y_i!

Two types of measurement error:

1. **Classical measurement error**: Random noise, independent of true value
   - $X_i^{\text{observed}} = X_i^{\text{true}} + \varepsilon_i$ where $\mathbb{E}[\varepsilon_i] = 0$, $\text{Cov}(X_i^{\text{true}}, \varepsilon_i) = 0$
2. **Non-classical measurement error**: Systematic, correlated with truth: $\text{Cov}(X_i^{\text{true}}, \varepsilon_i) \neq 0$
   - Example: Rich people overreport income, poor people underreport
   - Can bias estimates in *any* direction

Classical measurement error: Attenuation bias

Setup: True model: $Y_i = \alpha + \beta X_i^{\text{true}} + u_i$; observe $X_i^{\text{obs}} = X_i^{\text{true}} + \varepsilon_i$

Classical measurement error: Attenuation bias

Setup: True model: $Y_i = \alpha + \beta X_i^{\text{true}} + u_i$; observe $X_i^{\text{obs}} = X_i^{\text{true}} + \varepsilon_i$

Regressing Y_i on X_i^{obs}:

Classical measurement error: Attenuation bias

Setup: True model: $Y_i = \alpha + \beta X_i^{\text{true}} + u_i$; observe $X_i^{\text{obs}} = X_i^{\text{true}} + \varepsilon_i$

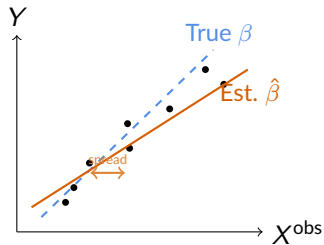
Regressing Y_i on X_i^{obs}:

$$\hat{\beta} = \frac{\text{Cov}(Y_i, X_i^{\text{obs}})}{\text{Var}(X_i^{\text{obs}})} = \frac{\beta \text{Var}(X_i^{\text{true}})}{\text{Var}(X_i^{\text{true}}) + \text{Var}(\varepsilon_i)} = \beta \cdot \underbrace{\frac{\text{Var}(X^{\text{true}})}{\text{Var}(X^{\text{true}}) + \text{Var}(\varepsilon)}}_{\text{attenuation factor } <1} \tag{17}$$

$\Longrightarrow$ **Attenuation bias**: Estimates biased toward zero

Visualizing measurement error

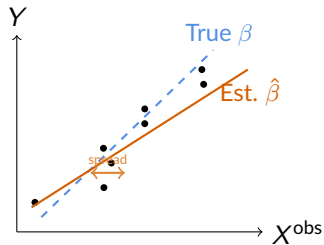
Noise in X (attenuation)



Noise **spreads out** X (horizontally)
$\Longrightarrow$ Flatter slope, bias toward zero

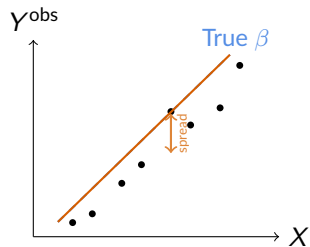
Visualizing measurement error

Noise in X (attenuation)



Noise **spreads out** X (horizontally)
$\Longrightarrow$ Flatter slope, bias toward zero

Noise in Y (no bias)



Noise **spreads out** Y **conditional on** X (vertically)
$\Longrightarrow$ Same slope, no bias, just less precision

Sources of bias: Self-reported data (Part 1)

Self-reports are common but problematic. Key sources of bias:

1. Recall bias

- People misremember past events/behaviors
- **Example**: Arthi et al. (2018, Tanzania): People report 4$\times$ more farm work hours when recalling vs. real-time measurement
- $\implies$ Large understatement of agricultural labor productivity

Sources of bias: Self-reported data (Part 1)

Self-reports are common but problematic. Key sources of bias:

1. Recall bias

- People misremember past events/behaviors
- **Example**: Arthi et al. (2018, Tanzania): People report 4$\times$ more farm work hours when recalling vs. real-time measurement
- $\Longrightarrow$ Large understatement of agricultural labor productivity

2. Social desirability bias / Experimenter demand

- Respondents give "socially acceptable" answers
- Want to please researchers or conform to norms
- Can measure with Crowne-Marlowe social desirability scale
- Adjust responses or use indirect questioning methods

Sources of bias: Self-reported data (Part 1)

Self-reports are common but problematic. Key sources of bias:

1. Recall bias

- People misremember past events/behaviors
- **Example**: Arthi et al. (2018, Tanzania): People report 4$\times$ more farm work hours when recalling vs. real-time measurement
- $\Longrightarrow$ Large understatement of agricultural labor productivity

2. Social desirability bias / Experimenter demand

- Respondents give "socially acceptable" answers
- Want to please researchers or conform to norms
- Can measure with Crowne-Marlowe social desirability scale
- Adjust responses or use indirect questioning methods

Sources of bias: Self-reported data (Part 2)

3. Question order effects

- Earlier questions influence responses to later ones
- Priming, anchoring, consistency effects

Sources of bias: Self-reported data (Part 2)

3. Question order effects

- Earlier questions influence responses to later ones
- Priming, anchoring, consistency effects

4. Framing effects

- Same information, different presentation $\implies$ different responses
- **Examples**:
  - "90% survival rate" vs. "10% mortality rate"
  - "Tax relief" vs. "tax cuts" vs. "tax decreases"

Sources of bias: Self-reported data (Part 3)

Sources of bias: Self-reported data (Part 3)

5. Selection bias

- Who answers the survey?
- Non-response often non-random
- **Example**: US political polling often has $<$5% response rate (Pew)
- Those who respond differ systematically from non-responders

Sources of bias: Self-reported data (Part 3)

5. Selection bias

- Who answers the survey?
- Non-response often non-random
- **Example**: US political polling often has $<$5% response rate (Pew)
- Those who respond differ systematically from non-responders

6. Hypothetical vs. incentivized responses

- Hypothetical: "Would you buy this product at $10?"
- Incentivized: Actually buying with real money
- **Recommendation**: Use incentive-compatible elicitation when possible
- Revealed preferences $>$ stated preferences

Partial solution to classical error: Creating indices

Idea: Combine multiple noisy measures to reduce measurement error

Simple case: Average two measures X_1 and X_2 of same construct: $X^{\text{index}} = (X_1 + X_2)/2$

If both have independent noise: Var(noise in index) $= \frac{1}{2}$Var(noise in each)

Partial solution to classical error: Creating indices

Idea: Combine multiple noisy measures to reduce measurement error

Simple case: Average two measures X_1 and X_2 of same construct: $X^{\text{index}} = (X_1 + X_2)/2$

If both have independent noise: Var(noise in index) $= \frac{1}{2}$Var(noise in each)

Derivation: Let $X_j = X^{\text{true}} + \varepsilon_j$ with $\varepsilon_1 \perp \varepsilon_2$. Then: $X^{\text{index}} = X^{\text{true}} + \frac{\varepsilon_1+\varepsilon_2}{2}$ and
Var $\left(\frac{\varepsilon_1+\varepsilon_2}{2}\right) = \frac{1}{4} \cdot 2\sigma_\varepsilon^2 = \frac{1}{2}\sigma_\varepsilon^2$

Partial solution to classical error: Creating indices

Idea: Combine multiple noisy measures to reduce measurement error

Simple case: Average two measures X_1 and X_2 of same construct: $X^{\text{index}} = (X_1 + X_2)/2$

If both have independent noise: Var(noise in index) $= \frac{1}{2}$Var(noise in each)

Derivation: Let $X_j = X^{\text{true}} + \varepsilon_j$ with $\varepsilon_1 \perp \varepsilon_2$. Then: $X^{\text{index}} = X^{\text{true}} + \frac{\varepsilon_1+\varepsilon_2}{2}$ and
$\text{Var}\left(\frac{\varepsilon_1+\varepsilon_2}{2}\right) = \frac{1}{4} \cdot 2\sigma_\varepsilon^2 = \frac{1}{2}\sigma_\varepsilon^2$

More generally: n measures $\implies$ noise variance reduced by factor of $1/n$

Partial solution to classical error: Creating indices

Idea: Combine multiple noisy measures to reduce measurement error

Simple case: Average two measures X_1 and X_2 of same construct: $X^{\text{index}} = (X_1 + X_2)/2$

If both have independent noise: Var(noise in index) $= \frac{1}{2}$Var(noise in each)

Derivation: Let $X_j = X^{\text{true}} + \varepsilon_j$ with $\varepsilon_1 \perp \varepsilon_2$. Then: $X^{\text{index}} = X^{\text{true}} + \frac{\varepsilon_1+\varepsilon_2}{2}$ and Var $\left(\frac{\varepsilon_1+\varepsilon_2}{2}\right) = \frac{1}{4} \cdot 2\sigma_\varepsilon^2 = \frac{1}{2}\sigma_\varepsilon^2$

More generally: n measures $\implies$ noise variance reduced by factor of $1/n$

Key assumption: Measurement errors independent (if correlated, gains smaller)

Weighted indices

Can do better than simple averages by weighting measures differently:

1. **Sum of standardized scores** (Kling et al. 2007)
   - Convert to z-scores $(X_j - \bar{X}_j)/\text{SD}(X_j)$, then average

Weighted indices

Can do better than simple averages by weighting measures differently:

1. **Sum of standardized scores** (Kling et al. 2007)
   - Convert to z-scores $(X_j - \bar{X}_j)/\text{SD}(X_j)$, then average

2. **Factor analysis**
   - Extract common factor; variables uncorrelated with others = mostly noise $\implies$ weight less

Weighted indices

Can do better than simple averages by weighting measures differently:

1. Sum of standardized scores (Kling et al. 2007)

- Convert to z-scores $(X_j - \bar{X}_j)/\text{SD}(X_j)$, then average

2. Factor analysis

- Extract common factor; variables uncorrelated with others = mostly noise $\implies$ weight less

3. Inverse-covariance weighting

- Variables uncorrelated with others provide independent info $\implies$ weight more
- Opposite logic to factor analysis—context determines which is better

Dealing with outliers and missing values

Outliers: Extreme values can drive results

Solutions:

- **Winsorization**: Cap extreme values at percentile (e.g., 1st/99th)
- **Trimming**: Drop extreme observations entirely
- **Transformation**: Log, inverse hyperbolic sine (for zeros)
- **Robust regression**: Median regression (LAD), Huber M-estimator

Dealing with outliers and missing values

Outliers: Extreme values can drive results

Solutions:

- **Winsorization**: Cap extreme values at percentile (e.g., 1st/99th)
- **Trimming**: Drop extreme observations entirely
- **Transformation**: Log, inverse hyperbolic sine (for zeros)
- **Robust regression**: Median regression (LAD), Huber M-estimator

Missing values: Often not missing at random

Solutions:

- **Imputation**: Replace with median, mean, or predicted values
- **Bounds / IPW**: As with attrition (Lee bounds, inverse probability weighting)
- **Missing indicator**: Include dummy for missing, set value to 0
- **Leave as missing**: Report sample size, acknowledge limitation

Best practices: No single “right” answer

Reality: For many measurement issues, no obviously “best” approach

The strategy:

1. **Be clear and justify your choices**
   - Explain why you chose specific method
   - Acknowledge limitations and alternatives

Best practices: No single “right” answer

Reality: For many measurement issues, no obviously “best” approach

The strategy:

1. **Be clear and justify your choices**
   - Explain why you chose specific method
   - Acknowledge limitations and alternatives
2. **Show robustness to alternative approaches**
   - Report results under different measurement choices
   - Include in main text or appendix
   - If results fragile to measurement choices, that’s important!

Best practices: No single “right” answer

Reality: For many measurement issues, no obviously “best” approach

The strategy:

1. **Be clear and justify your choices**
   - Explain why you chose specific method
   - Acknowledge limitations and alternatives
2. **Show robustness to alternative approaches**
   - Report results under different measurement choices
   - Include in main text or appendix
   - If results fragile to measurement choices, that’s important!

Remember: Good measurement is as important as good identification

Can have perfect RCT but meaningless results if measuring wrong thing

Key Takeaways

Main lessons from today:

1. **Choosing controls**: Not all controls are good—think causally (DAGs). Bad controls can bias estimates or block mechanisms.

Key Takeaways

Main lessons from today:

1. **Choosing controls**: Not all controls are good—think causally (DAGs). Bad controls can bias estimates or block mechanisms.
2. **Heterogeneous effects**: Average effects may mask important variation. Use modern ML methods (GATES, causal forests) to study heterogeneity rigorously.

Key Takeaways

Main lessons from today:

1. **Choosing controls**: Not all controls are good—think causally (DAGs). Bad controls can bias estimates or block mechanisms.
2. **Heterogeneous effects**: Average effects may mask important variation. Use modern ML methods (GATES, causal forests) to study heterogeneity rigorously.
3. **Attrition**: Can create selection bias even in RCTs. Test for differential attrition and use bounds (Lee, Manski) or IPW.

Key Takeaways

Main lessons from today:

1. **Choosing controls**: Not all controls are good—think causally (DAGs). Bad controls can bias estimates or block mechanisms.
2. **Heterogeneous effects**: Average effects may mask important variation. Use modern ML methods (GATES, causal forests) to study heterogeneity rigorously.
3. **Attrition**: Can create selection bias even in RCTs. Test for differential attrition and use bounds (Lee, Manski) or IPW.
4. **Multiple hypothesis testing**: Testing many hypotheses inflates false positives. Control FDR (Benjamini-Hochberg) or FWER (Bonferroni). Pre-specify analyses.

Key Takeaways

Main lessons from today:

1. **Choosing controls**: Not all controls are good—think causally (DAGs). Bad controls can bias estimates or block mechanisms.
2. **Heterogeneous effects**: Average effects may mask important variation. Use modern ML methods (GATES, causal forests) to study heterogeneity rigorously.
3. **Attrition**: Can create selection bias even in RCTs. Test for differential attrition and use bounds (Lee, Manski) or IPW.
4. **Multiple hypothesis testing**: Testing many hypotheses inflates false positives. Control FDR (Benjamini-Hochberg) or FWER (Bonferroni). Pre-specify analyses.
5. **Replication**: Results less robust than we thought. Researcher degrees of freedom, publication bias, and winner's curse all contribute. Transparency and robustness checks are essential.

Key Takeaways (cont.)

6. **Standard errors**: Always use heteroskedasticity-robust SEs. Cluster when observations correlated within groups (panel, spatial, etc.). Most common mistake in applied work.

Key Takeaways (cont.)

6. **Standard errors**: Always use heteroskedasticity-robust SEs. Cluster when observations correlated within groups (panel, spatial, etc.). Most common mistake in applied work.
7. **Spillovers**: SUTVA violations matter. Design two-step clustered RCTs to measure indirect effects. Use recentering approach (Borusyak & Hull) to test for spillovers.

Key Takeaways (cont.)

6. **Standard errors**: Always use heteroskedasticity-robust SEs. Cluster when observations correlated within groups (panel, spatial, etc.). Most common mistake in applied work.
7. **Spillovers**: SUTVA violations matter. Design two-step clustered RCTs to measure indirect effects. Use recentering approach (Borusyak & Hull) to test for spillovers.
8. **Measurement error**: Classical error in X causes attenuation bias. Self-reports have many biases (recall, social desirability, framing). Combine multiple measures to reduce noise.

Key Takeaways (cont.)

6. **Standard errors**: Always use heteroskedasticity-robust SEs. Cluster when observations correlated within groups (panel, spatial, etc.). Most common mistake in applied work.
7. **Spillovers**: SUTVA violations matter. Design two-step clustered RCTs to measure indirect effects. Use recentering approach (Borusyak & Hull) to test for spillovers.
8. **Measurement error**: Classical error in X causes attenuation bias. Self-reports have many biases (recall, social desirability, framing). Combine multiple measures to reduce noise.

Overall theme: Even with perfect identification, empirical work requires careful attention to practical challenges. Good research design anticipates and addresses these issues proactively.

GATES/CLAN: Why sample splitting matters

Overfitting problem: Using *same* data to train ML and estimate GATES $\implies$ biased estimates

Solution: Sample splitting

- **Auxiliary sample**: Train ML predictor
- **Main sample**: Estimate GATES/CLAN (not used to choose groups)

Aggregation: Results may depend on the particular random split

- Repeat with many splits, report **median** estimates
- More robust than any single split

Implementation: R package `GenericML`

Back