Lecture 03:
Randomized controlled trials I

PPHA 34600
Prof. Fiona Burlig

Harris School of Public Policy
University of Chicago

From last time: selection is an issue

Recall that there are lots of things we want to estimate.

We need to get around selection bias to do this.

In other words, we need:

$$E[Y_i(1)] = E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$$

and

$$E[Y_i(0)] = E[Y_i(0)|D_i = 0] = E[Y_i(0)|D_i = 1]$$

Regression equivalent:

$$E[\varepsilon_i|D_i] = 0$$

Random assignment as a solution

When treatment status is randomly assigned,

$$F(X, \varepsilon | D = 1) = F(X, \varepsilon | D = 0) = F(X, \varepsilon)$$

In words:

The distribution of **both** observables (Xs) **and** unobservables (εs) is the same for treated and untreated units!

There is **no selection problem** by construction!

Again, but mathier

When D, treatment, is **randomly assigned**:

- D is independent of $Y(0)$ and $Y(1)$
- The distribution of $Y_i(0)|D_i$ is equal to the unconditional distribution
- The distribution of $Y_i(1)|D_i$ is equal to the unconditional distribution
- $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
- $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

Again, but mathier

When D, treatment, is **randomly assigned**:

- D is independent of $Y(0)$ and $Y(1)$
- The distribution of $Y_i(0)|D_i$ is equal to the unconditional distribution
- The distribution of $Y_i(1)|D_i$ is equal to the unconditional distribution
- $E[Y_i(1)|D_i = 1] = E[Y_i(1)]$
- $E[Y_i(0)|D_i = 0] = E[Y_i(0)]$

As a result:

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

$$= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

$$= E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

This bears repeating

Under randomization:

$$\tau^{ATE} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

This bears repeating

Under randomization:

$$\tau^{ATE} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

We can easily estimate this from data:

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

We can estimate the ATE simply from the difference in means between treated and "control" group.

This bears repeating

Under randomization:

$$\tau^{ATE} = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

We can easily estimate this from data:

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

We can estimate the ATE simply from the difference in means between treated and “control” group.

Obvious (?) caveat: We still can't get τ_i, because we only observe i once.

Evaluating an RCT

This is not a class on how to do RCTs

- As always, the devil is in the details
- Field experiments are *hard!*
- But supposing you've got one...

Evaluating an RCT

This is not a class on how to do RCTs

- As always, the devil is in the details
- Field experiments are *hard!*
- But supposing you've got one...

Basic RCT checklist

- ☐ Verify random assignment
- ☐ Check compliance with treatment
- ☐ Estimate the ATE (or other things...)

What is this experiment trying to learn?

When running an RCT, you want to have a “research question” in mind:

What is the causal effect of [program x] on [outcome y]?

What is this experiment trying to learn?

When running an RCT, you want to have a "research question" in mind:

What is the causal effect of [program x] on [outcome y]?

Why do we need an RCT to study this?

What is this experiment trying to learn?

When running an RCT, you want to have a "research question" in mind:

What is the causal effect of [program x] on [outcome y]?

Why do we need an RCT to study this?

- Program X targets certain individuals
- Individuals who choose to participate look different than non-participants
- Others?

Understanding RCTs

Basic ingredients for an RCT:

- What is the research design?
  - What is the unit of randomization?
  - How was randomization performed?
- What are the outcomes of interest?

Verifying random assignment

Did randomization "work"?

- Randomization should mean treated and control units are similar
- This is true *in expectation*, not necessarily for one draw

Verifying random assignment

Did randomization "work"?

- Randomization should mean treated and control units are similar
- This is true *in expectation*, not necessarily for one draw

Testing whether randomization was effective:

- We want T and C to be similar on observables **and** unobservables
- We can only test this for observables
- To check this, we "test for balance":
- Compare mean outcomes for T vs. C *at baseline* (before treatment) or in fixed characteristics

  $\rightarrow$ Implementation: Regress $Y_i^{baseline} = \alpha + \tau D_i + \nu_i$

Checking for balance

Three things to check for:

1. Did they test for all outcome variables?
2. Are differences statistically significant?
3. Are magnitudes economically meaningful?

Checking compliance with treatment

Did assignment to treatment affect treatment status?

Trying to verify whether...

- Units assigned to treatment were actually treated
- Units assigned to control were *not* treated

There is often substantial non-compliance. We'll talk more about exactly how to deal with this issue next time.

Thinking about non-compliance

We will treat this more formally next time

For now, non-compliance changes the interpretation of our estimates:

> Rather than asking "What does treatment do to our outcome activities?"...
>
> ... we're asking "What does offering treatment do to our outcome?"

This may be the policy-relevant quantity

We want to estimate the ATE

Recall that the ATE is just:

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

Since we have random assignment, we can estimate this as:

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

We want to estimate the ATE

Recall that the ATE is just:

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

Since we have random assignment, we can estimate this as:

$$\hat{\tau}^{ATE} = \overline{Y(1)} - \overline{Y(0)}$$

Regression is a convenient way to do this:

$$Y_i = \alpha + \tau D_i + \varepsilon_i$$

Since our $E[\varepsilon|D_i] = 0$ assumption is satisfied (why?), $\hat{\tau} = \hat{\tau}^{ATE}$

Estimating treatment effects

We'll often see things that look like this:

$$y_{ia} = \alpha + \tau Treat_{ia} + \gamma \mathbf{X}_a^{\text{baseline}} + \varepsilon_{ia}$$

where:

- y_{ia} are outcomes for household i in area a
- α is a constant
- $Treat_{ia}$ is a treatment dummy (think D_i)
- $\mathbf{X}_a^{\text{baseline}}$ is a set of baseline area controls
- ε_{ia} is an error term

What is this equation estimating?

$$y_{ia} = \alpha + \tau \, Treat_{ia} + \gamma \mathbf{X}_a^{\text{baseline}} + \varepsilon_{ia}$$

This differs from our basic regression a bit:

- There's an *i and* an *a*
- We have $\gamma \mathbf{X}_a^{\text{baseline}}$

Let's unpack each of these in turn...

Randomization by area, data on individuals

We have *i*-ndividual level data, but *a*-rea level randomization

Randomizing at a higher level of aggregation is common:

- Some questions can't be answered at *i* level (no personal bank branches)
- Ethics concerns: can sometimes delay implementation for a whole group; hard for individuals
- Reduce spillovers (more on this later)

Randomization by area, data on individuals

We have *i*-ndividual level data, but *a*-rea level randomization

Randomizing at a higher level of aggregation is common:

- Some questions can't be answered at *i* level (no personal bank branches)
- Ethics concerns: can sometimes delay implementation for a whole group; hard for individuals
- Reduce spillovers (more on this later)

Randomizing at a higher level affects the analysis:

- Interpretation is different (what exactly is treatment?)
- Getting standard errors right requires either:
  1. Estimate *i*-level effects, but cluster at *a*-level
     or
  2. Averaging outcomes at the group level (weight by individuals per group)

Adding controls

If D_i is randomly assigned, we don't need X_i!

<u>We often add controls anyway:</u>

Adding controls

If D_i is randomly assigned, we don't need X_i!

We often add controls anyway:

- Controlling for X_i should not affect $\hat{\tau}$
  - $\rightarrow$ Why?

Adding controls

If D_i is randomly assigned, we don't need X_i!

We often add controls anyway:

- Controlling for X_i should not affect $\hat{\tau}$

  $\rightarrow$ Why?

- Controlling for X_i will affect the standard error on $\hat{\tau}$

  $\rightarrow$ Why?

Adding controls

If D_i is randomly assigned, we don't need X_i!

We often add controls anyway:

- Controlling for X_i should not affect $\hat{\tau}$
  - $\rightarrow$ Why?
- Controlling for X_i will affect the standard error on $\hat{\tau}$
  - $\rightarrow$ Why?

☠ do **not** control for post-treatment outcomes

Adding bad controls

First rule of RCT club:

- Do **not** control for post-treatment outcomes
- Do **not** control for post-treatment outcomes

→ If treatment affects these outcomes, you can get bias!

Simple example:

- Suppose microfinance impacts business ownership
- By random assignment, households with and without loans have the same potential income
- Once we condition on business ownership, this is no longer true!

We can use simulated data to think about this

Type of household	Potential business ownership: Without MF	Potential business ownership: With MF	Potential income: Without MF	Potential income: With MF	Average earnings by ownership: Without MF	Average earnings by ownership: With MF
Never owner	No	No	1,000	1,500		
Moved by MF	No	Yes	2,000	2,500		
Always owner	Yes	Yes	3,000	3,500		

We can use simulated data to think about this

Type of household	Potential business ownership: Without MF	Potential business ownership: With MF	Potential income: Without MF	Potential income: With MF	Average earnings by ownership: Without MF	Average earnings by ownership: With MF
Never owner	No	No	1,000	1,500	Don't own: 1,500	Don't own: 1,500
Moved by MF	No	Yes	2,000	2,500		Own: 3,000
Always owner	Yes	Yes	3,000	3,500	Own: 3,000	

We can use simulated data to think about this

Type of household	Potential business ownership		Potential income		Average earnings by ownership	
	Without MF	With MF	Without MF	With MF	Without MF	With MF
Never owner	No	No	1,000	1,500	Don't own: 1,500	Don't own: 1,500
Moved by MF	No	Yes	2,000	2,500		Own: 3,000
Always owner	Yes	Yes	3,000	3,500	Own: 3,000	

- The return to MFI is 500 for everyone...
- But once we condition on ownership, it looks like the return is 0!

→ This is because we don't have random assignment **within** ownership!

We can use simulated data to think about this

<table>
<tr><th rowspan="2">Type of household</th><th colspan="2">Potential business ownership</th><th colspan="2">Potential income</th><th colspan="2">Average earnings by ownership</th></tr>
<tr><th>Without MF</th><th>With MF</th><th>Without MF</th><th>With MF</th><th>Without MF</th><th>With MF</th></tr>
<tr><td>Never owner</td><td>No</td><td>No</td><td>1,000</td><td>1,500</td><td rowspan="2">Don’t own: 1,500</td><td>Don’t own: 1,500</td></tr>
<tr><td>Moved by MF</td><td>No</td><td>Yes</td><td>2,000</td><td>2,500</td><td rowspan="2">Own: 3,000</td></tr>
<tr><td>Always owner</td><td>Yes</td><td>Yes</td><td>3,000</td><td>3,500</td><td>Own: 3,000</td></tr>
</table>

- The return to MFI is 500 for everyone...
- But once we condition on ownership, it looks like the return is 0!

$\rightarrow$ This is because we don’t have random assignment **within** ownership!

Do not control for post-treatment outcomes!

We can also estimate heterogeneous effects

Heterogeneous effects are straightforward:

$$\tau(X_1 = x_1) = E[Y_i(1)|X_1 = x_1] - E[Y_i(0)|X_1 = x_1]$$

We can also estimate heterogeneous effects

Heterogeneous effects are straightforward:

$$\tau(X_1 = x_1) = E[Y_i(1)|X_1 = x_1] - E[Y_i(0)|X_1 = x_1]$$

We typically estimate these in two ways:

1. Add an **interaction term** to the regression:

$$y_i = \alpha + \tau \, Treat_i + \gamma \, Treat_i \cdot X_i + \delta X_i + \varepsilon_i$$

→ Make sure to add both the interaction and the base term

We can also estimate heterogeneous effects

Heterogeneous effects are straightforward:

$$\tau(X_1 = x_1) = E[Y_i(1)|X_1 = x_1] - E[Y_i(0)|X_1 = x_1]$$

We typically estimate these in two ways:

1. Add an **interaction term** to the regression:

$$y_i = \alpha + \tau Treat_i + \gamma Treat_i \cdot X_i + \delta X_i + \varepsilon_i$$

   → Make sure to add both the interaction and the base term

2. Estimate the regression **separately** by heterogeneity
   → Equivalent to a *fully* interacted model

Estimate heterogeneity by pre-determined characteristics only!

A note on assumptions for the RCT

We still need several assumptions for the RCT to work:

- $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
  and
  $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$
  - $\rightarrow$ We "get this" via randomization, but only in expectation

A note on assumptions for the RCT

We still need several assumptions for the RCT to work:

- $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
  and
  $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$

  $\rightarrow$ We "get this" via randomization, but only in expectation

- Perfect compliance

  $\rightarrow$ Kinda. More on this next class

A note on assumptions for the RCT

We still need several assumptions for the RCT to work:

- $E[Y_i(1)|D_i = 1] = E[Y_i(1)|D_i = 0]$
  and
  $E[Y_i(0)|D_i = 1] = E[Y_i(0)|D_i = 0]$

  $\rightarrow$ We "get this" via randomization, but only in expectation

- Perfect compliance

  $\rightarrow$ Kinda. More on this next class

- No spillovers: "SUTVA"

  - Stable Unit Treatment Value Assumption: D_i doesn't affect j's potential outcomes

  $\rightarrow$ Kinda. More on this in two classes

Application: Tuition-free college in Michigan

Dynarski, Libassi, Michelmore, and Owen (2020 NBER WP)

Policy challenge:

- Gaps in educational attainment between low- and high-income kids
- These persist among high-achieving kids
- College has big labor market impacts

Intervention:

- Promise of free tuition and fees to UMichigan-Ann Arbor
- Information sent to students ("HAIL" scholarship)

Tuition-free college in Michigan: The experiment

→ **Lesson for you as MPPs:** RCTs are doable in high-stakes contexts!

This is a **group-level randomization design**:

- Population: high-achieving, low-income seniors in public school
- Randomization done at the *school* level (why?)
- Schools stratified by HAIL-eligible populations

Outcomes of interest

Outcome data is administrative information (!)

Outcomes:

- Application to UMich
- Admission to UMich
- Enrollment at UMich
- Other college choices

Other characteristics:

- Student demographics
- Student performance, incl GPA and test scores

Balance?

Characteristic	Mean		P-value
	Control schools	Treated schools	
Region, urbanicity, and distance			
Upper Peninsula	0.150	0.130	0.344
	(0.016)	(0.015)	
West Central	0.449	0.476	0.359
	(0.022)	(0.022)	
Southeast	0.401	0.394	0.788
	(0.021)	(0.022)	
Suburban	0.340	0.360	0.537
	(0.021)	(0.021)	
City	0.129	0.100	0.148
	(0.015)	(0.013)	
Rural	0.530	0.540	0.718
	(0.022)	(0.022)	
Distance from UM	93.2	96.4	0.529
	(3.545)	(3.673)	

Balance?

Student academic characteristics			
Average SAT (or equivalent)	1254	1260	0.194
	(2.690)	(2.896)	
Average GPA	3.823	3.833	0.208
	(0.006)	(0.006)	
Proportion limited English proficient	0.002	0.004	0.410
	(0.001)	(0.001)	
Proportion receiving special education services	0.009	0.013	0.367
	(0.003)	(0.004)	
Proportion who sent ACT/SAT scores to UM	0.365	0.377	0.695
	(0.015)	(0.016)	
UM application rate in 2015	0.067	0.055	0.016
	(0.004)	(0.004)	
Missing 2015 UM application rate	0.004	0.020	0.015
	(0.003)	(0.006)	

Balance?

Characteristic	Mean		P-value
	Control schools	Treated schools	
School size			
# of 11th grade students in school	189.1	175.1	0.055
	(0.003)	(0.006)	
# of HAIL students in school	3.8	3.9	0.649
	(0.140)	(0.163)	
F-test for joint significance: p-value			0.0004
Number of schools	526	500	1,026
Number of students	1,978	1,932	3,910

Compliance?

Since we observe the outcomes for all students, and therefore all schools, there is no attrition due to non-response. We do not observe whether a student actually receives the information packet (i.e. is effectively treated), and students assigned to the control group cannot be treated, so we do not adjust for non-compliance.

Regression specification and parameters of interest

These authors very simply estimate (modified for our notation):

$$Y_j = \alpha + \tau D_j + \gamma S_j + \beta \mathbf{X}_j + \varepsilon_j$$

where:

- Y_j is our outcome for school j
- D_j is our treatment indicator
- S_j is a stratum fixed effect
- $\mathbf{X}_j$ are controls
- ε_j is an error term

<u>Note:</u> Data are collected at the individual level, but collapsed to the school-(cohort) level

Findings

Outcome	Treatment effect		Control mean
Applied	0.416	0.413	0.259
	(0.021)	(0.019)	
Admitted	0.174	0.163	0.149
	(0.019)	(0.017)	
Enrolled	0.149	0.141	0.117
	(0.018)	(0.016)	
Strata dummies	X	X	
Covariates		X	
Number of schools	1,026		
Number of students	3,910		

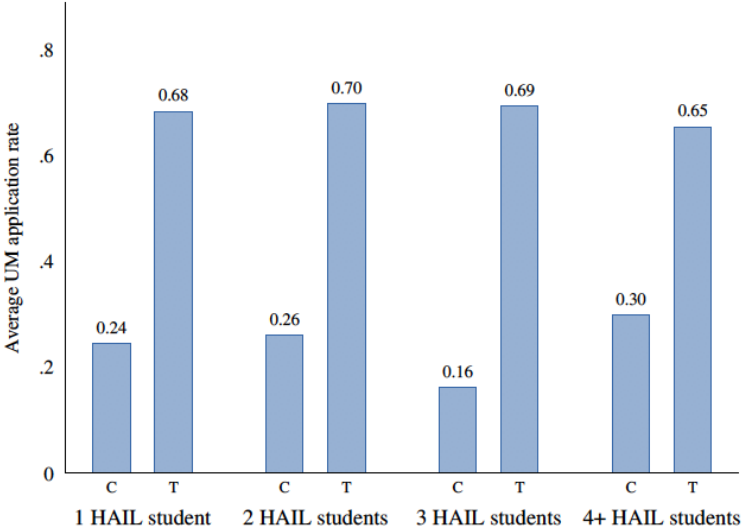
Findings

College attended	Treatment effect	Control mean
Highly competitive or above	0.146	0.135
	(0.018)	
UM	0.146	0.107
	(0.016)	
Highly competitive or above other than UM	0.000	0.028
	(0.007)	
Four-year	0.074	0.675
	(0.020)	
Two-year	-0.035	0.116
	(0.013)	
Any	0.039	0.791
	(0.018)	
In Michigan	0.045	0.727
	(0.020)	
Public in Michigan	0.062	0.645
	(0.021)	
Outside Michigan	-0.006	0.064
	(0.010)	
Number of schools	1,026	
Number of students	3,910	

Findings

College attended	Attended fall following high school graduation		Attended two consecutive falls following high school graduation	
	Treatment effect	Control mean	Treatment effect	Control mean
Highly competitive or above	0.153 (0.024)	0.129	0.135 (0.023)	0.126
UM	0.147 (0.022)	0.104	0.128 (0.022)	0.102
Four-year	0.091 (0.028)	0.651	0.109 (0.029)	0.557
Any	0.057 (0.025)	0.779	0.079 (0.027)	0.683
Number of schools		529		
Number of students		2,108		

Heterogeneity



Recap

TL;DR:

1. RCTs are great!
2. Experiments solve our selection problem
3. Be very careful with adding controls