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.

<u>Obvious (?) caveat:</u> 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!

We often add controls anyway:

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 $\hat{\tau}$
  - → 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!

<u>Simple example:</u>

- 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		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		
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		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	

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!

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

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	

- 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!

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

  → We "get this" via randomization, but only in expectation

- Perfect compliance

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

  → We "get this" via randomization, but only in expectation

- Perfect compliance

  → Kinda. More on this next class

- No spillovers: "SUTVA"

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

  → Kinda. More on this in two classes

Application: Audits of polluting firms

Duflo, Greenstone, Pande, and Ryan (QJE 2013)

Policy challenge:

- Pollution from industrial plants is very high in Gujarat
- Auditors responsible for monitoring are paid by the polluting firms (!)

Intervention:

- Firms pay into an independent account
- Auditors are randomly assigned to firms
- Some firms were visited for back-checks

Pollution audits in Gujarat: The experiment

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

This is a **stratified randomization design**:

- Sample: 633 high-polluting plants
- Stratification on region
- 50% of firms were randomized into treatment group
- Ineligible plants eliminated after randomization
- 20% of plant readings got back-checks

Outcomes of interest

Outcome data measured throughout 2009-10 and at endline

Outcomes of interest:

- Pollution levels → regulatory compliance
- Pollution levels relative to back-checks ("truth-telling")

Balance?

	(1) Treatment	(2) Control	(3) Difference
Panel A: Plant characteristics			
Capital investment INR 50 m to 100 m (= 1)	0.092	0.14	−0.051
	[0.29]	[0.35]	(0.033)
Located in industrial estate (= 1)	0.57	0.53	0.042
	[0.50]	[0.50]	(0.051)
Textiles (= 1)	0.88	0.93	−0.030
	[0.33]	[0.26]	(0.025)
Effluent to common treatment (= 1)	0.41	0.35	0.078
	[0.49]	[0.48]	(0.049)
Wastewater generated (kl/day)	420.5	394.6	35.4
	[315.9]	[323.4]	(31.6)
Lignite used as fuel (= 1)	0.71	0.77	−0.024
	[0.45]	[0.42]	(0.029)
Diesel used as fuel (= 1)	0.29	0.25	0.038
	[0.45]	[0.43]	(0.046)
Air emissions from flue gas (= 1)	0.85	0.87	−0.0095
	[0.35]	[0.33]	(0.016)
Air emissions from boiler (= 1)	0.93	0.92	0.026
	[0.26]	[0.27]	(0.027)
Bag filter installed (= 1)	0.24	0.34	−0.10**
	[0.43]	[0.47]	(0.046)
Cyclone installed (= 1)	0.087	0.079	0.0010
	[0.28]	[0.27]	(0.027)
Scrubber installed (= 1)	0.41	0.41	−0.018
	[0.49]	[0.49]	(0.050)

Balance?

Panel B: Regulatory interactions in year prior to study

Whether audit submitted (= 1)	0.82	0.81	0.022
	[0.38]	[0.39]	(0.038)
Any equipment mandated (= 1)	0.42	0.49	−0.047
	[0.50]	[0.50]	(0.047)
Any inspection conducted (= 1)	0.79	0.78	0.016
	[0.41]	[0.42]	(0.042)
Any citation issued (= 1)	0.28	0.24	0.035
	[0.45]	[0.43]	(0.045)
Any water citation issued (= 1)	0.12	0.12	−0.0031
	[0.33]	[0.33]	(0.034)
Any air citation issued (= 1)	0.027	0.0052	0.021*
	[0.16]	[0.072]	(0.013)
Any utility disconnection (= 1)	0.098	0.094	0.0029
	[0.30]	[0.29]	(0.031)
Any bank guarantee posted (= 1)	0.033	0.026	0.0045
	[0.18]	[0.16]	(0.017)

Compliance?

Noncompliance not an issue here:

> Overall, we collected 2,953 pollution samples from 408 plants in the study sample, an average of 7.2 pollutants per plant.[15] Attrition in the endline survey was balanced across treatment and control groups.[16]

Regression specification and parameters of interest

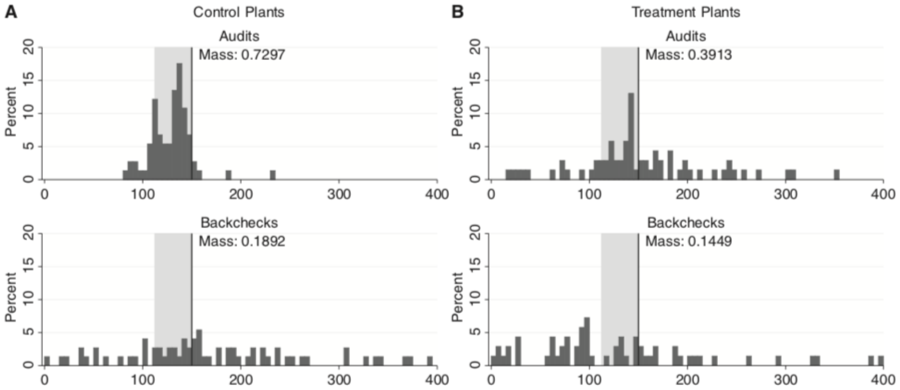
These authors estimate (a slightly more complicated version of):

$$y_{ir} = \alpha + \tau D_{ir} + \alpha_r + \varepsilon_{ir}$$

where:

- y_{ir} is the outcome for firm i in region r
- α is a constant
- D_{ir} is a treatment indicator
- α_r is a fixed effect for region
- ε_{ir} is an error term

Findings



A
Control Plants
Audits
Mass: 0.7297
Percent
0 5 10 15 20
0 100 200 300 400
Backchecks
Mass: 0.1892
Percent
0 5 10 15 20
0 100 200 300 400
B
Treatment Plants
Audits
Mass: 0.3913
Percent
0 5 10 15 20
0 100 200 300 400
Backchecks
Mass: 0.1449
Percent
0 5 10 15 20
0 100 200 300 400

Findings

ENDLINE POLLUTANT CONCENTRATIONS ON TREATMENT STATUS

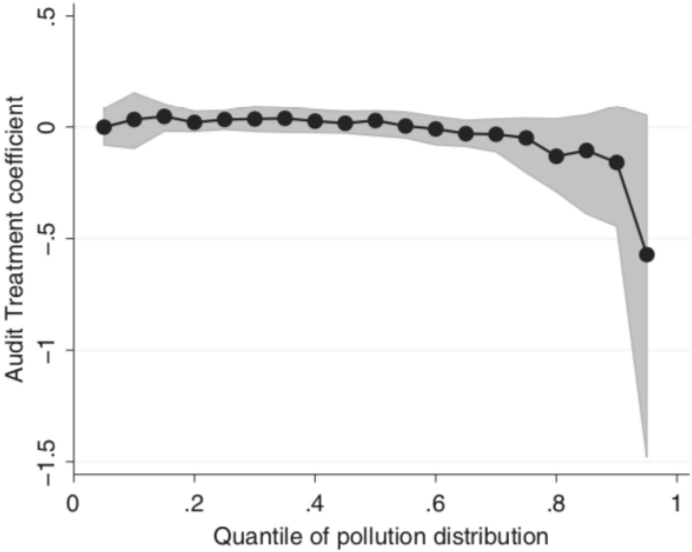
	(1) All pollutants	(2) Water pollutants	(3) Air pollutants
Panel A: Dependent variable: Level of pollutant in endline survey, all pollutants (standard deviations relative to backcheck mean)			
Audit treatment assigned (= 1)	−0.211**	−0.300*	−0.053
	(0.099)	(0.159)	(0.057)
Control mean	0.076	0.114	0.022
Observations	1439	860	579
Panel B: Dependent variable: Compliance (dummy for pollutant in endline survey at or below regulatory standard)			
Audit treatment assigned (=1)	0.027	0.039	0.002
	(0.027)	(0.039)	(0.028)
Control mean	0.573	0.516	0.656
Observations	1,439	860	579

Findings

Compliance in Audits Relative to Backchecks by Treatment Status

	(1) All pollutants	(2) Water pollutants	(3) Air pollutants
Panel A: Dependent variable: Narrow compliance (dummy for pollutant between 75% and 100% of regulatory standard)			
Audit report × Treatment group	−0.185***	−0.212***	−0.143***
	(0.034)	(0.044)	(0.046)
Audit report (= 1)	0.270***	0.297***	0.230***
	(0.025)	(0.034)	(0.033)
Treatment group (= 1)	−0.0034	−0.013	0.011
	(0.0176)	(0.025)	(0.024)
Control mean in backchecks	0.097	0.110	0.077
Panel B: Dependent variable: Compliance (dummy for pollutant at or below regulatory standard)			
Audit report × Treatment group	−0.234***	−0.166***	−0.345***
	(0.039)	(0.050)	(0.056)
Audit report (= 1)	0.288***	0.273***	0.311***
	(0.023)	(0.033)	(0.032)
Treatment group (= 1)	0.058*	0.0075	0.145***
	(0.034)	(0.0477)	(0.041)
Control mean in backchecks	0.557	0.538	0.586
Observations	2236	1378	858

Heterogeneity



Recap

TL;DR:

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