Lecture 06: Evaluation of evaluations

PPHA 34600
Prof. Fiona Burlig

Harris School of Public Policy
University of Chicago

From last time: RCTs are the bee's knees

Randomized controlled trials are super powerful:

- Random assignment allows us to solve our selection problem
- We can implement them with tweaks to handle challenges:
  - *Noncompliance:* Dividing τ^{ITT} by share of compliers $\rightarrow \tau^{LATE}$
  - *Spillovers:* Proper design to avoid or measure
- More opportunities than you might imagine for implementation

Moving out of RCT land

We will spend the rest of the course on other research designs:

- Randomized controlled trials (RCTs)
- Trying to control for observable things
- Panel data
- Instrumental variables
- Regression discontinuity
- Big Data and machine learning

Why leave RCT land?

RCTs are the gold standard for a reason, but:

- They can be expensive
- Some programs require evaluation at scale
- RCTs can't always be implemented
- **There's a lot to learn from non-RCTs**

The ideal experiment

Even as we move away from RCTs, it's useful to consider the **ideal experiment**

The ideal experiment

Even as we move away from RCTs, it's useful to consider the **ideal experiment**

- "What experiment would I run to answer this question?"
- Useful to nail down your question of interest
- Valuable to think through problems with your non-RCT

The ideal experiment

Even as we move away from RCTs, it's useful to consider the **ideal experiment**

- "What experiment would I run to answer this question?"
- Can be totally feasible (RED for energy efficiency upgrades)...
- ...or totally infeasible (randomly warm one Earth while keeping the other cold)

Testing quasi-experimental methods

We want to know how well our non-RCT toolkit works...
How can we test this?

Testing quasi-experimental methods

We want to know how well our non-RCT toolkit works...
How can we test this?

→ Compare an actual RCT to other methods *on the same data*

- Requires a setting with an RCT
- (Usually) toss the control group
- Use quasi-experimental methods to estimate $\hat{\tau}$
- Compare the RCT to the quasi-experimental methods

Testing quasi-experimental methods

We want to know how well our non-RCT toolkit works...
How can we test this?

→ Compare an actual RCT to other methods *on the same data*

- Requires a setting with an RCT
- (Usually) toss the control group
- Use quasi-experimental methods to estimate $\hat{\tau}$
- Compare the RCT to the quasi-experimental methods

→ What do we learn?

Testing quasi-experimental methods

We want to know how well our non-RCT toolkit works...
How can we test this?

→ Compare an actual RCT to other methods *on the same data*

- Requires a setting with an RCT
- (Usually) toss the control group
- Use quasi-experimental methods to estimate $\hat{\tau}$
- Compare the RCT to the quasi-experimental methods

→ What do we learn?

This is often referred to as a "LaLonde exercise" after LaLonde (1986)

Testing quasi-experimental methods

Comparing RCTs with quasi-experiments teaches us:

- If the estimates are the same:

Testing quasi-experimental methods

Comparing RCTs with quasi-experiments teaches us:

- If the estimates are the same:
  - Good sign!
  - This suggests that the quasi-experimental method is working properly

Testing quasi-experimental methods

Comparing RCTs with quasi-experiments teaches us:

- If the estimates are the same:
  - Good sign!
  - This suggests that the quasi-experimental method is working properly
- If the estimates are different:

Testing quasi-experimental methods

Comparing RCTs with quasi-experiments teaches us:

- If the estimates are the same:
  - Good sign!
  - This suggests that the quasi-experimental method is working properly
- If the estimates are different:
  - ☠☠☠
  - Something is likely going wrong with the non-RCT

Testing quasi-experimental methods

Comparing RCTs with quasi-experiments teaches us:

- If the estimates are the same:
  - Good sign!
  - This suggests that the quasi-experimental method is working properly
- If the estimates are different:
  - ☠☠☠
  - Something is likely going wrong with the non-RCT

Context is really important for this!

Leveraging an RCT we know and love

Blast from the past: We'll use the SMUD pricing RCT

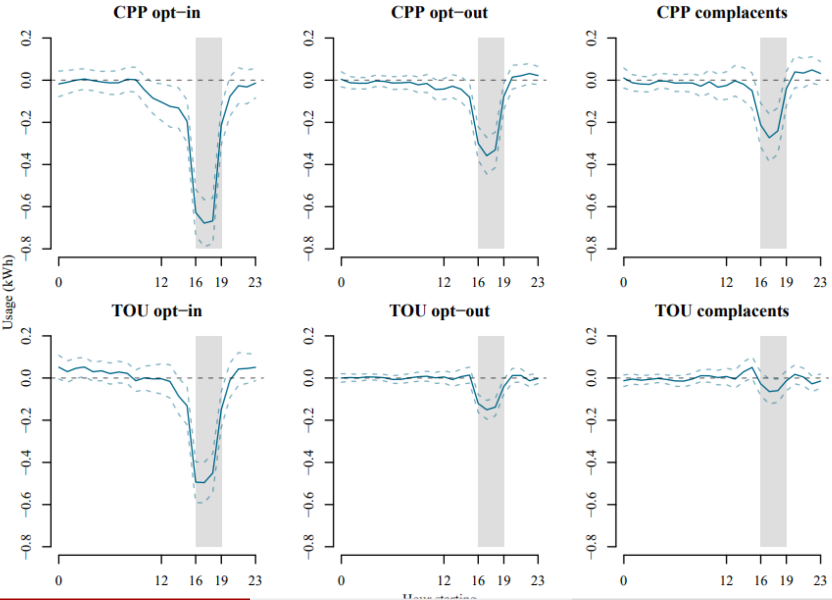
Policy issue:

- The cost of providing electricity is time-varying
- Prices typically aren't
- This causes large welfare losses

Program:

- SMUD (randomly) implemented time-varying pricing
- Experimental run: 2011-2013
- Two flavors: "time-of-use" (TOU) and "critical peak pricing" (CPP)
- Both opt-in and opt-out versions

Fowlie & Wolfram et al results recap



Why did we want an RCT in this context?

Research question:
What is the effect of time-varying electricity pricing on consumption?

Why did we want an RCT in this context?

Research question:

What is the effect of time-varying electricity pricing on consumption?

Potential for selection into treatment on characteristics

→ People who participate in utility programs look different

→ Suppose *no response* to treatment: If units who choose treatment have different consumption patterns, we'll mistakenly measure $\hat{\tau} > 0$

Why did we want an RCT in this context?

Research question:
What is the effect of time-varying electricity pricing on consumption?

Potential for selection into treatment on characteristics

- $\rightarrow$ People who participate in utility programs look different
- $\rightarrow$ Suppose *no response* to treatment: If units who choose treatment have different consumption patterns, we'll mistakenly measure $\hat{\tau} > 0$

Potential for selection into treatment on τ_i

- $\rightarrow$ People who choose to get treated may have different price sensitivity
- $\rightarrow$ We know some of this is happening! (two LATEs)

What would the naive estimator do?

Recall our naive estimator:

$$\tau^N = \bar{Y}(1) - \bar{Y}(0)$$

What would the naive estimator do?

Recall our naive estimator:

$$\tau^N = \bar{Y}(1) - \bar{Y}(0)$$

Why is this problematic for electricity pricing?

Going beyond the naive estimator

We have a good idea that the naive estimator won't work...

... so we turn to alternative **research designs**

Going beyond the naive estimator

We have a good idea that the naive estimator won't work...

... so we turn to alternative **research designs**

A research design:

- Tries to solve the selection problem without randomization
- Invokes stronger assumptions than the RCT
- Allows us to make progress without randomization
- Best-case scenario: mimics an RCT

Spurlock et al test three common methods

We will cover each of these methods in detail later:

Spurlock et al test three common methods

We will cover each of these methods in detail later:

1. **Difference-in-differences**
   - Essentially compares treated and untreated units over time
   - Intuition: I am similar to myself, treated or not...
   - ... but to rule out other stuff happening, I use never-treated units

Spurlock et al test three common methods

We will cover each of these methods in detail later:

1. **Difference-in-differences**
   - Essentially compares treated and untreated units over time
   - Intuition: I am similar to myself, treated or not...
   - ... but to rule out other stuff happening, I use never-treated units
2. **(Propensity score) matching**
   - Essentially a sophisticated way of "controlling for stuff"
   - Tries to generate a (non-experimental) control group
   - Goal is to make this similar to the treated group

Spurlock et al test three common methods

We will cover each of these methods in detail later:

1. **Difference-in-differences**
   - Essentially compares treated and untreated units over time
   - Intuition: I am similar to myself, treated or not...
   - ... but to rule out other stuff happening, I use never-treated units
2. **(Propensity score) matching**
   - Essentially a sophisticated way of "controlling for stuff"
   - Tries to generate a (non-experimental) control group
   - Goal is to make this similar to the treated group
3. **Regression discontinuity**
   - Essentially compares just-treated units to just-untreated units
   - Leverages cutoffs in policy

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

1. Compare treated unit i in time t to (pre-treatment) i in time $t - 1$
   - Intuitive part: compare me to myself pre/post treatment

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

1. Compare treated unit i in time t to (pre-treatment) i in time $t-1$
   - Intuitive part: compare me to myself pre/post treatment
2. Compare untreated unit j in time t to untreated j in time $t-1$
   - Why do this?

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

1. Compare treated unit i in time t to (pre-treatment) i in time $t-1$
   - Intuitive part: compare me to myself pre/post treatment
2. Compare untreated unit j in time t to untreated j in time $t-1$
   - Why do this?

   $\rightarrow$ Control for common shocks to everyone
3. Subtract difference (1) from difference (2):

$$y_{it} = \alpha + \tau D_{it} + \underbrace{\gamma_i}_{i \text{ to itself}} + \underbrace{\delta_t}_{i \text{ to } j \text{ over time}} + \varepsilon_{it}$$

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

1. Compare treated unit i in time t to (pre-treatment) i in time $t-1$
   - Intuitive part: compare me to myself pre/post treatment
2. Compare untreated unit j in time t to untreated j in time $t-1$
   - Why do this?
   - $\rightarrow$ Control for common shocks to everyone
3. Subtract difference (1) from difference (2):

$$y_{it} = \alpha + \tau D_{it} + \underbrace{\gamma_i}_{i \text{ to itself}} + \underbrace{\delta_t}_{i \text{ to } j \text{ over time}} + \varepsilon_{it}$$

For this to work, we require:

- Consumption for treated units is trending similarly to untreated units

Difference in difference approach

The DD approach compares units to themselves over time (3 steps):

1. Compare treated unit i in time t to (pre-treatment) i in time $t-1$
   - Intuitive part: compare me to myself pre/post treatment
2. Compare untreated unit j in time t to untreated j in time $t-1$
   - Why do this?

   → Control for common shocks to everyone
3. Subtract difference (1) from difference (2):

$$y_{it} = \alpha + \tau D_{it} + \underbrace{\gamma_i}_{i \text{ to itself}} + \underbrace{\delta_t}_{i \text{ to } j \text{ over time}} + \varepsilon_{it}$$

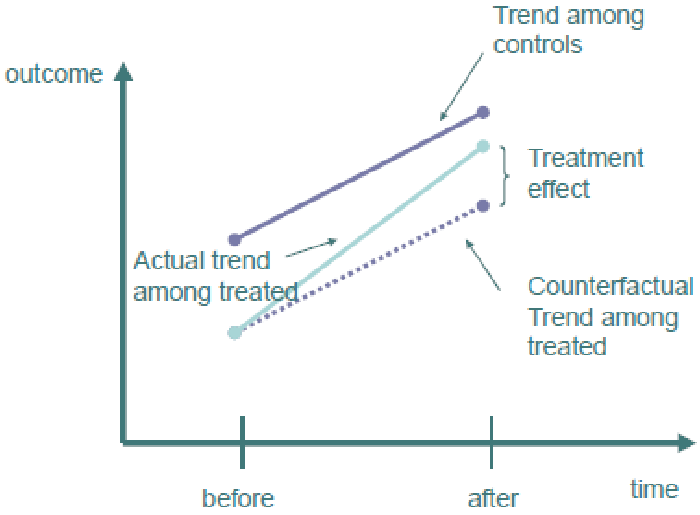
For this to work, we require:

- Consumption for treated units is trending similarly to untreated units

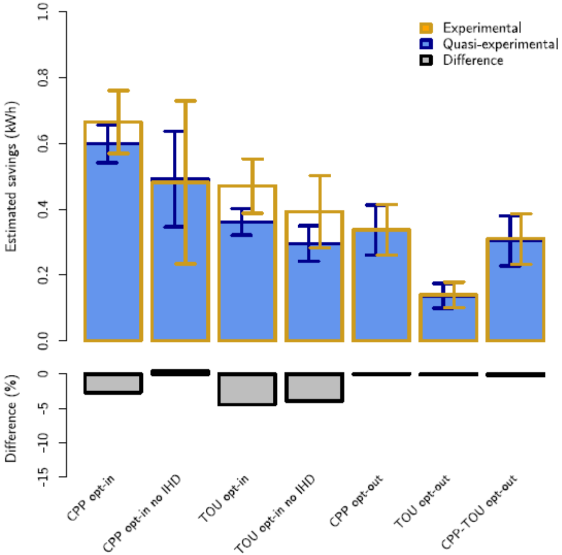
Note: Spurlock et al drop encouraged-but-untreated units

→ Was this necessary?

Difference in difference intuition



Comparing experimental and diff-in-diff results



Matching approach

(Propensity score) matching is a fancy way to control for stuff (2 steps):

Matching approach

(Propensity score) matching is a fancy way to control for stuff (2 steps):

1. Use pre-treatment consumption to find untreated units that look like treated units
    - Eliminate dis-similar untreated units from the sample

Matching approach

(Propensity score) matching is a fancy way to control for stuff (2 steps):

1. Use pre-treatment consumption to find untreated units that look like treated units
   - Eliminate dis-similar untreated units from the sample
2. Estimate treatment effects for all treated units and selected untreated units only:

$$y_{it} = \alpha + \tau D_{it} + \gamma_i + \delta_t + \varepsilon_{it}$$

   - (You can do this in several different ways)

Matching approach

(Propensity score) matching is a fancy way to control for stuff (2 steps):

1. Use pre-treatment consumption to find untreated units that look like treated units
   - Eliminate dis-similar untreated units from the sample
2. Estimate treatment effects for all treated units and selected untreated units only:

$$y_{it} = \alpha + \tau D_{it} + \gamma_i + \delta_t + \varepsilon_{it}$$

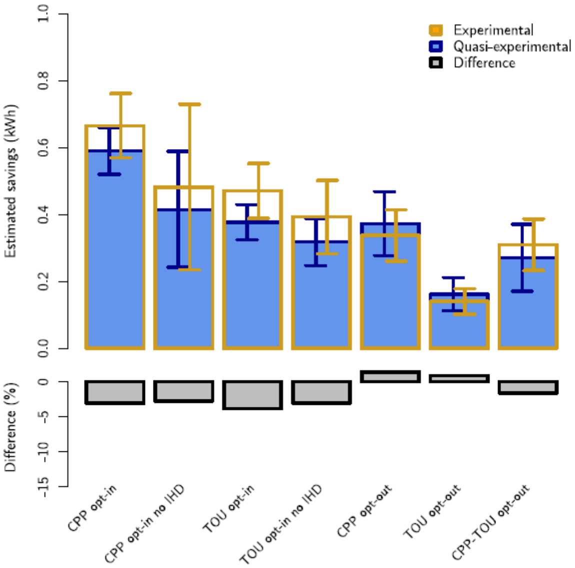
   - (You can do this in several different ways)

For this to work, we require:

- Our selection control soaks up everything that matters!

Note: Same approach as DD, but now controlling for more

Comparing experimental and propensity score results



Regression discontinuity approach

The RD compares unit i to a nearly-identical unit j below a cutoff:

Regression discontinuity approach

The RD compares unit i to a nearly-identical unit j below a cutoff:

- Use a policy cutoff to "randomize" treatment
  - Compare a unit with pre-period consumption just below 100 to a unit with pre-period consumption just above 100

Regression discontinuity approach

The RD compares unit i to a nearly-identical unit j below a cutoff:

- Use a policy cutoff to “randomize” treatment
  - Compare a unit with pre-period consumption just below 100 to a unit with pre-period consumption just above 100

For this to work, we require:

- Units on either side of the cutoff are otherwise similar

Regression discontinuity approach

The RD compares unit i to a nearly-identical unit j below a cutoff:

- Use a policy cutoff to "randomize" treatment
  - Compare a unit with pre-period consumption just below 100 to a unit with pre-period consumption just above 100

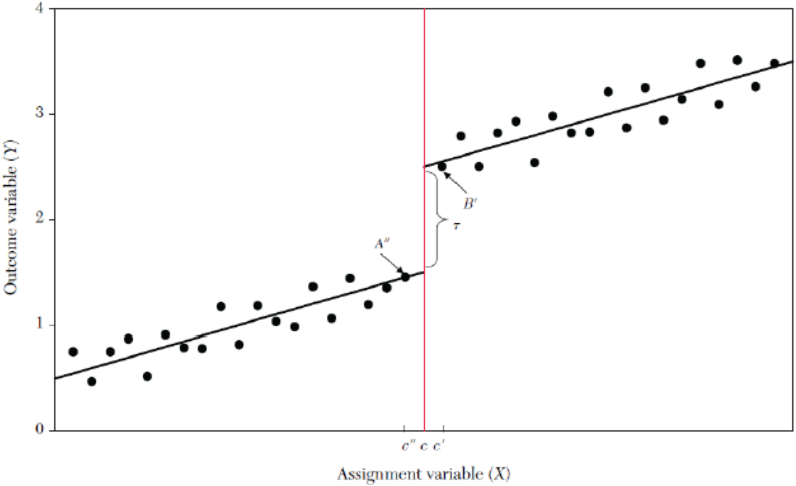
For this to work, we require:

- Units on either side of the cutoff are otherwise similar

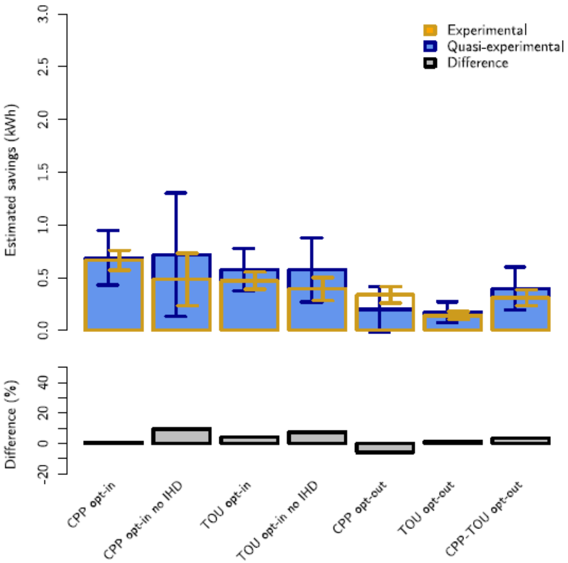
Note: Spurlock et al construct fake cutoffs

- They stitch together control group units with treatment group units
- The stitching point is their artificial cutoff

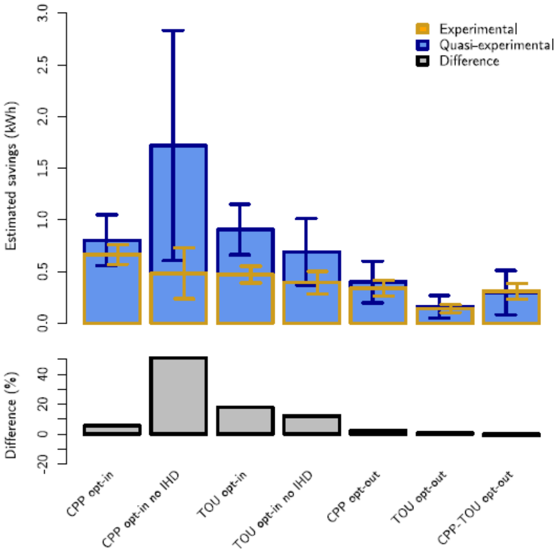
Regression discontinuity intuition



Comparing experimental and regression discontinuity



Comparing experimental and regression discontinuity



Punchlines

Spurlock et al find:

- Difference in difference estimates *understate* treatment effects
  - Due to unabsorbed selection

Punchlines

Spurlock et al find:

- Difference in difference estimates *understate* treatment effects
  - Due to unabsorbed selection
- Propensity score estimates *understate* treatment effects
  - Due to unabsorbed selection (controls make it worse!)

Punchlines

Spurlock et al find:

- Difference in difference estimates *understate* treatment effects
  - Due to unabsorbed selection
- Propensity score estimates *understate* treatment effects
  - Due to unabsorbed selection (controls make it worse!)
- Regression discontinuity estimates *overstate* treatment effects
  - Due to unabsorbed selection
  - OR estimating a different LATE

Punchlines

Spurlock et al find:

- Difference in difference estimates *understate* treatment effects
  - Due to unabsorbed selection
- Propensity score estimates *understate* treatment effects
  - Due to unabsorbed selection (controls make it worse!)
- Regression discontinuity estimates *overstate* treatment effects
  - Due to unabsorbed selection
  - OR estimating a different LATE
- Opt-out treatments are less biased than opt-in treatments
  - Intuition: We do better with a less-selected treatment

Exercise caution with non-experimental results

"Even though I was unable to evaluate all non-experimental methods, this evidence suggests that policymakers should be aware that the available non-experimental evaluations...may contain large and unknown biases resulting from specification errors." – LaLonde (1986)

Recap

TL;DR:

1. RCTs are (still) great!

2. Quasi-experimental methods can get things wrong

3. We don't usually have a good experimental benchmark (☠)