Lecture 01:
Introduction to Program Evaluation

PPHA 34600

Prof. Fiona Burlig

Harris School of Public Policy
University of Chicago

Why are we here?

I have two main goals for this course:

1. Introduce you to program evaluation, and build familiarity with modern methods

2. Prepare you to be a *consumer* of program evaluation

Why are we here?

I have two main goals for this course:

1. Introduce you to program evaluation, and build familiarity with modern methods

2. Prepare you to be a *consumer* of program evaluation

"Program evaluation" is a blanket term

Broadly, understanding whether a program "works" requires many steps:

1. Needs assessment: is this program necessary?
2. Theory of change: how is this program expected to work?
3. Process analysis: was the program implemented properly?
4. Impact evaluation: what did the program do?
5. Cost-benefit analysis: is this program cost-effective / efficient?

"Program evaluation" is a blanket term

Broadly, understanding whether a program "works" requires many steps:

1. Needs assessment: Is this program necessary?
2. Theory of change: How is this program expected to work?
3. Process analysis: Was the program implemented properly?
4. **Impact evaluation: What did the program do?**
5. Cost-benefit analysis: Is this program cost-effective / efficient?

We focus on impact evaluation: It's hard, requiring its own course...

... but we'll end up doing a little bit of the others as well

Program evaluation is critical for good public policy

We spend billions of dollars annually trying to do stuff.

Program evaluation lets us understand what happens as a result.

Program evaluation is critical for good public policy

We spend billions of dollars annually trying to do stuff.

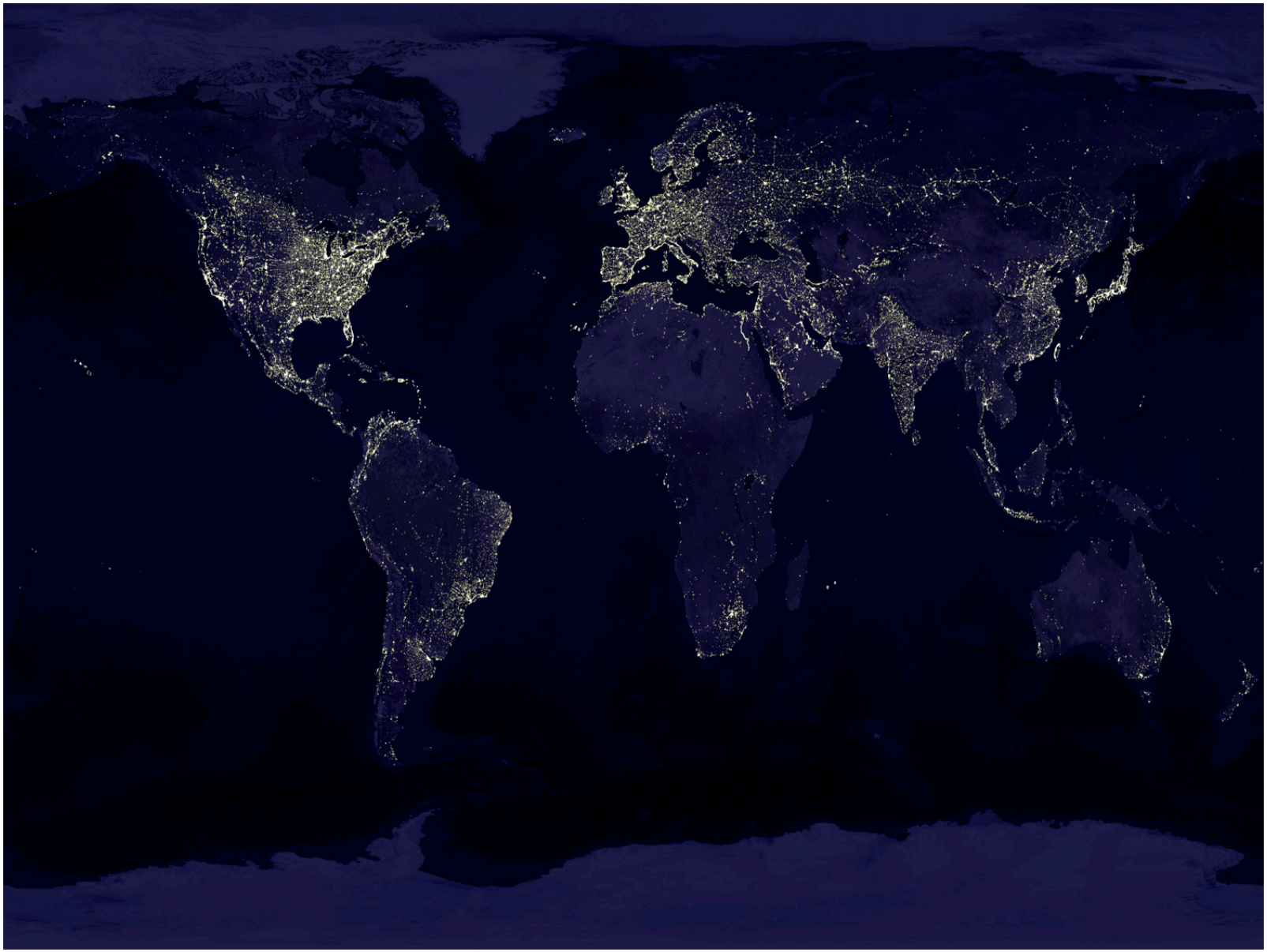
Program evaluation lets us understand what happens as a result.

The key objective is to *improve Welfare*.

Program evaluation enables us to:

- Understand whether a program is effective, ineffective, or other
- Diagnose program shortcomings
- Create better future programs as a result

A concrete example: rural electrification in India



A concrete example: rural electrification in India

Policy challenge: Over a billion people live without access to electricity

- Over 400 million of them live in India

A concrete example: rural electrification in India

Policy challenge: Over a billion people live without access to electricity

- Over 400 million of them live in India

Program: *Rajiv Gandhi Grameen Vidyutikaran Yojana* national electrification policy introduced in 2005

- All unelectrified villages supposed to get power
- Brought transmission and distribution infrastructure to rural areas

What could RGGVY have done?

We begin by distinguishing **outcomes** from **impacts:**

Outcomes: Things that we could "potentially" observe

- Incomes
- Years of schooling
- Number of agricultural workers
- Small business operations

What could RGGVY have done?

We begin by distinguishing **outcomes** from **impacts:**

Outcomes: Things that we could "potentially" observe

- Incomes
- Years of schooling
- Number of agricultural workers
- Small business operations

Impacts: Changes in outcomes *caused* by the policy

- Changes in incomes *as a result* of RGGVY
- Changes in the number of years of schooling *caused by* electricity
- Changes in the agricultural work force *due to* access to power
- Changes in small businessses *stemming from* rural electrification

What could RGGVY have done?

We begin by distinguishing **outcomes** from **impacts:**

Outcomes: Things that we could "potentially" observe

- Incomes
- Years of schooling
- Number of agricultural workers
- Small business operations

Impacts: Changes in outcomes *caused* by the policy

- Changes in incomes *as a result* of RGGVY
- Changes in the number of years of schooling *caused by* electricity
- Changes in the agricultural work force *due to* access to power
- Changes in small businessses *stemming from* rural electrification

When we discuss changes, you should ask: "Compared to what?"

Was RGGVY effective?

We need to consider all possible outcomes (not just what we observe)

Realized outcomes:

- Outcomes that we *actually* observe

Potential outcomes:

- All possible outcomes we *could have* observed
- Spans both actual and alternative programs

Counterfactual outcomes:

- Outcomes that we *would have observed* without (with) the program

→ Exposed units: what would have happened without the program

→ Non-exposed units: what would have happened with the program

Was RGGVY effective?

We need to consider all possible outcomes (not just what we observe)

Realized outcomes:

- Outcomes that we *actually* observe

Potential outcomes:

- All possible outcomes we *could have* observed
- Spans both actual and alternative programs

Counterfactual outcomes:

- Outcomes that we *would have observed* without (with) the program

→ Exposed units: what would have happened without the program

→ Non-exposed units: what would have happened with the program

What is the outcome of RGGVY compared to the counterfactual?

Measuring impacts requires a formal definition

Define: $D_i \in \{0, 1\}$ as the **treatment indicator** for unit i

- When unit i is treated, $D_i = 1$
- When unit i is not treated, $D_i = 0$

Define: $Y_i(D_i)$ as the **outcome** for unit i as a function of D_i

- When unit i is treated, we observe $Y(1)$
- When unit i is not treated, we observe $Y(0)$

Then, the **impact** of treatment for unit i is just:

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

If we say "RGGVY improved incomes in rural India," we are implying: "A village that got RGGVY had higher incomes *relative to that same village* without RGGVY"

How can one unit be in two states at once?

This is the **fundamental problem of causal inference**:

We only ever get to observe $Y_i(1)$ **or** $Y_i(0)$, **but not both!**

How can one unit be in two states at once?

This is the **fundamental problem of causal inference**:

We only ever get to observe $Y_i(1)$ **or** $Y_i(0)$, **but not both!**

Remember when I said program evaluation was hard?

It's worse than that: it's **impossible** to observe individual-specific impacts

If we can't compare me to m[*i*]self, what do we do?

Rather than measuring impacts for **each individual**, we can estimate the...

Average Treatment Effect (ATE):

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

Intuitively: Even if I can't see $Y_i(1)$ and $Y_i(0)$ at the same time, I can see $Y(1)$ and $Y(0)$ *on average* at once.

One possible approach to estimating the ATE

We can try to estimate the ATE by computing the...

Naive estimator:

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

where $\overline{Y(D_i)}$ is the average Y for units with treatment status $D_i \in \{0, 1\}$.

NB: This is a *sample* average, as opposed to the population function $E[]$. We'll abuse this notation aggressively throughout this class.

Did RGGVY cause incomes to rise?

After the introduction of RGGVY, Indian politicians pointed out:

- "Power availability has helped many to get self-employed in avocations in which they have skills and this had led to increase in employment and also income." – RGGVY evaluation report
- Indeed, employment rates in the villages with electricity were about 5% higher than in villages without electricity.

Did RGGVY cause incomes to rise?

After the introduction of RGGVY, Indian politicians pointed out:

- "Power availability has helped many to get self-employed in avocations in which they have skills and this had led to increase in employment and also income." – RGGVY evaluation report
- Indeed, employment rates in the villages with electricity were about 5% higher than in villages without electricity.

How should we interpret these averages?

- Is the conclusion being drawn necessarily correct?
- What assumptions underly this conclusion?
- Why might these assumptions fail to hold?

What's going wrong?

When we use the naive estimator, we are assuming that:

$$E[Y_i(1)] = E[Y_i(1) \mid D_i = 1] = E[Y_i(1) \mid D_i = 0]$$

and

$$E[Y_i(0)] = E[Y_i(0) \mid D_i = 0] = E[Y_i(0) \mid D_i = 1]$$

That is, we assume that the *unconditional* expectation of Y is the same as the *conditional* expectation of Y.

Put differently: we assume that the average of units with $D_i = 0$ are a good counterfactual for units with $D_i = 1$.

This ignores **selection**: the idea that treated units and untreated units may differ, **even absent treatment**.

Unpacking the selection problem

The selection problem comes from that pesky FPCI:

- If we could observe $Y_i(1)$ *and* $Y_i(0)$, we'd have no issues.
- Instead, we see $Y_i(1)$ and $Y_j(0)$, where $i \neq j$.

We have to ask: Why did i get treated and j not?

Why did *i* get treated while *j* didn't?

Almost always: *i* and *j* are fundamentally different.

We typically classify **two types** of "differences":

Selection on observables:

- Treated and untreated units differ along lines we can see
- Ex: electrified villages are wealthier than unelectrified villages

Selection on unobservables:

- Treated and untreated units differ along lines we can't see
- Ex: electrified villages like electricity better than unelectrified villages

Note: Depending on your dataset, observables in one context may be unobservable in others

- Practically speaking, you'll always have unobservables

What does selection mean for our naive estimator?

We're trying to measure the effect of D on Y:

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

If all electrified villages belong to the ruling political party...

... what does this approach measuring?

What does selection mean for our naive estimator?

We're trying to measure the effect of D on Y:

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

If all electrified villages belong to the ruling political party...

... what does this approach measuring?

We can't distinguish between effects of electrification and politics!

What does selection mean for our naive estimator?

We're trying to measure the effect of D on Y:

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

If all electrified villages belong to the ruling political party...

... what does this approach measuring?

We can't distinguish between effects of electrification and politics!

(Remember how I said program evaluation was hard?)

How do we address selection?

A (slightly less) naive approach is **constructing bounds**

A popular approach is known as Manski bounds:

- **Idea:** think about best- and worst-case scenarios
- Requires only weak assumptions
- Gives some sense of how bad the selection problem might be

Constructing Manski bounds

Let's continue with our rural electrification example:

Indian data	
Pr(Has access to electricity)	0.11
Pr(Does not have electricity access)	0.89
E[Above poverty line]	0.61
E[Above poverty line | electricity access]	0.77
E[Above poverty line | no electricity access]	0.59
N	579,659

$$\tau^N = 0.77 - 0.59 = \mathbf{0.18}$$

$\rightarrow$ electrified villages are 18 percentage points more likely to be *above* the poverty line!

Constructing Manski bounds

We already know that τ^N might not be equal to τ^{ATE}.

To construct Manski bounds:

$$\begin{aligned} E[Y_i(1)] &= Pr(D_i = 1)E[Y_i(1)|D_i = 1] + Pr(D_i = 0)E[Y_i(1)|D_i = 0] \\ &= 0.11 \times 0.77 + 0.89 \times \mathbf{E[Y_i(1)|D_i = 0]} \end{aligned}$$

and

$$\begin{aligned} E[Y_i(0)] &= Pr(D_i = 1)E[Y_i(0)|D_i = 1] + Pr(D_i = 0)E[Y_i(0)|D_i = 0] \\ &= 0.11 \times \mathbf{E[Y_i(0)|D_i = 1]} + 0.89 \times 0.59 \end{aligned}$$

Constructing Manski bounds

Using the fact that $0 \leq Y_i \leq 1$, we can bound $E[Y_i(1)|D_i = 0]$:

$$E[Y_i(1)]^{\text{Upper Bound}} = 0.11 \times 0.77 + 0.89 \times 1 = 0.9747$$

$$E[Y_i(1)]^{\text{Lower Bound}} = 0.11 \times 0.77 + 0.89 \times 0 = 0.0847$$

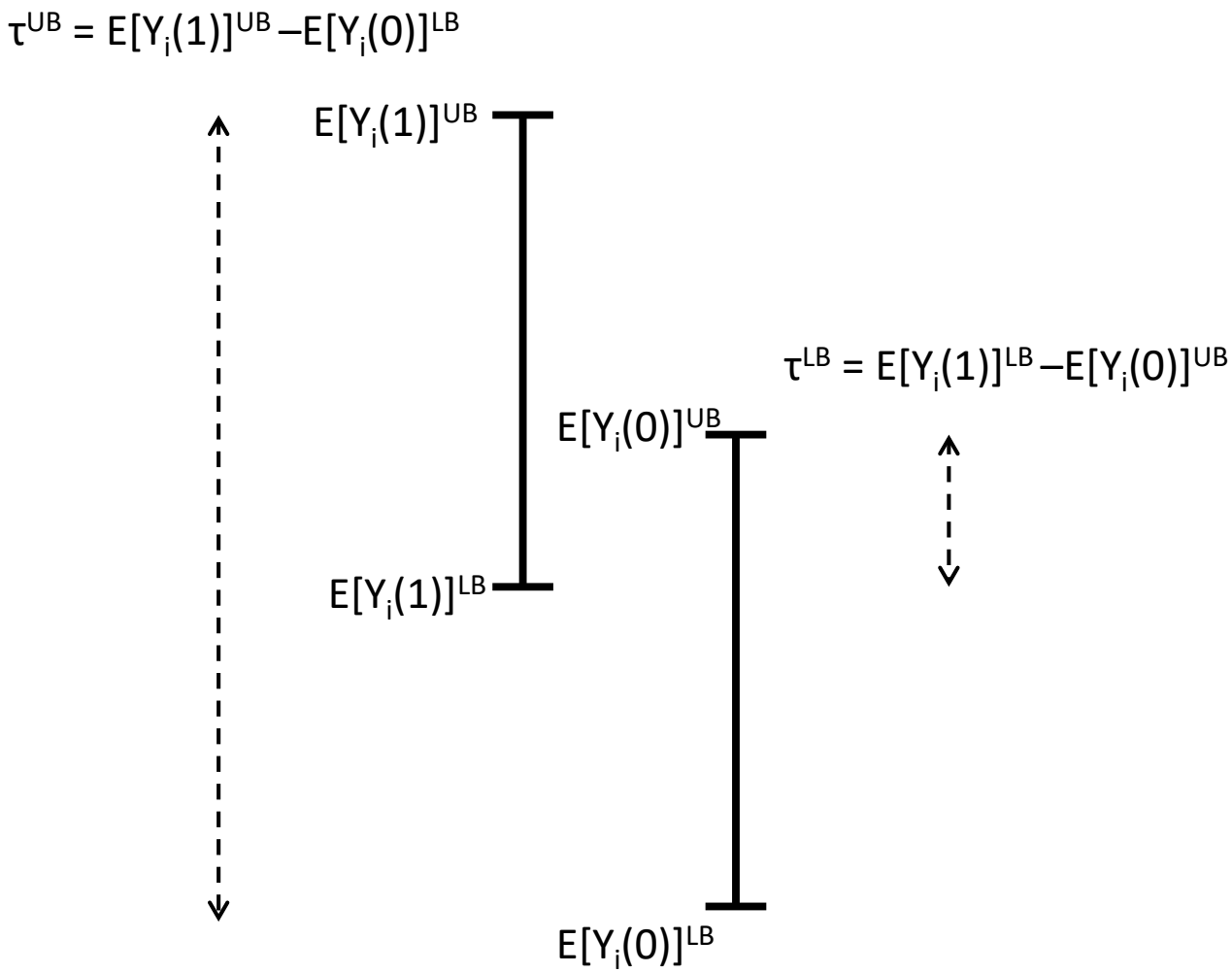
Similarly for $E[Y_i(0)|D_i = 1]$:

$$E[Y_i(0)]^{\text{Upper Bound}} = 0.11 \times 1 + 0.89 \times 0.59 = 0.6351$$

$$E[Y_i(0)]^{\text{Lower Bound}} = 0.11 \times 0 + 0.89 \times 0.59 = 0.5251$$

$\rightarrow$ **We can put these together to bound the ATE**

(Graphically) bounding the ATE



Bounding the ATE

The **upper bound** will be:

$$\tau^{\text{Upper Bound}} = E[Y_i(1)]^{\text{Upper Bound}} - E[Y_i(0)]^{\text{Lower Bound}}$$
$$\tau^{\text{Upper Bound}} = 0.9747 - 0.5251 = 0.4496$$

The **lower bound** will be:

$$\tau^{\text{Lower Bound}} = E[Y_i(1)]^{\text{Lower Bound}} - E[Y_i(0)]^{\text{Upper Bound}}$$
$$\tau^{\text{Upper Bound}} = 0.0847 - 0.6351 = -0.5504$$

The Manski bounds are (approximately) [-0.55, 0.45].
You could drive a bus through this!

No, really, how do we address selection?

This is the subject of the remainder of the course

We use research designs:

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

Why are we here?

I have two main goals for this course:

1. Introduce you to program evaluation, and build familiarity with modern methods
2. Prepare you to be a *consumer* of program evaluation

What does it mean to be a good consumer?

Not all impact evaluations are created equal

What does it mean to be a good consumer?

Not all impact evaluations are created equal

Female hurricanes are deadlier than male hurricanes

Kiju Jung[a,1], Sharon Shavitt[a,b,1], Madhu Viswanathan[a,c], and Joseph M. Hilbe[d]

[a]Department of Business Administration and [b]Department of Psychology, Institute of Communications Research, and Survey Research Laboratory, and [c]Women and Gender in Global Perspectives, University of Illinois at Urbana–Champaign, Champaign, IL 61820; and [d]Department of Statistics, T. Denny Sanford School of Social and Family Dynamics, Arizona State University, Tempe, AZ 85287-3701

Edited* by Susan T. Fiske, Princeton University, Princeton, NJ, and approved May 14, 2014 (received for review February 13, 2014)

Do people judge hurricane risks in the context of gender-based expectations? We use more than six decades of death rates from US hurricanes to show that feminine-named hurricanes cause significantly more deaths than do masculine-named hurricanes. Laboratory experiments indicate that this is because hurricane names lead to gender-based expectations about severity and this, in turn, guides respondents' preparedness to take protective action. This finding indicates an unfortunate and unintended consequence of the gendered naming of hurricanes, with important implications for policymakers, media practitioners, and the general public concerning hurricane communication and preparedness.

The goal is to help you tell the difference!

Recap

TL;DR:

1. Program evaluation is important (and hard!)
2. Selection bias is a big issue

For next class

Topics:

- Parameters of interest
- Regression as our primary tool

Reading: Angrist and Pischke

- *Mastering 'metrics:* pp. 82-97
- *Mostly Harmless Econometrics:* pp 27-64