The Fundamental Problem of Econometrics

EC 320: Introduction to Econometrics

Emmett Saulnier
Spring 2022

Housekeeping

Analytical problem set 1 due tomorrow (4/8) at 11:59pm, it covers the statistics review.

Computational problem set 2 is due on Monday (4/11) at 11:59pm, it is all about data wrangling in the Tidyverse, which you talked about in lab yesterday.

I am busy during my usual Friday office hours this week --- so I will have to change them to be **Today (4/7) from 3-4pm**. Feel free to email me on Friday with questions about the problem set.

Prologue

Statistics Inform Policy

Policy: In 2017, the University of Oregon started requiring first-year students to live on campus.

Rationale: First-year students who live on campus fare better than those who live off campus.

- *80 percent more likely* to graduate in four years.
- Second-year retention rate *5 percentage points higher*.
- GPAs *0.13 points higher*, on average.

Do these comparisons suggest that the policy will improve student outcomes?

Do they describe the effect of living on campus?

Do they describe ***something else*?**

Other Things Equal

The UO's interpretation of those comparisons warrants skepticism.

- The decision to live on campus is probably related to family wealth and interest in school.
- Family wealth and interest in school are also related to academic achievement.

Why? The difference in outcomes between those on and off campus is not an ***all else equal***[*] comparison.

Upshot: We can't attribute the difference in outcomes solely to living on campus.

[*] *All else equal = ceteris paribus, all else held constant, etc.*

Other Things Equal

A high bar

When all other factors are held constant, statistical comparisons detect causal relationships.

(Micro)economics has developed a comparative advantage in understanding where ***all else equal*** comparisons can and cannot be made.

- Anyone can retort "*correlation doesn't necessarily imply causation*."
- Understanding *why* is difficult, but useful for learning from data.

The Fundamental Problem of Econometrics

Causal Identification

Goal

Identify the effect of a **treatment**, D_i, on individual i's **outcome**, $Y_{D,i}$.

Ideal data

Ideally, we could calculate the **treatment effect** *for each individual* i as

$$Y_{1,i} - Y_{0,i}$$

- $Y_{1,i}$ is the outcome for person i when she receives the treatment.
- $Y_{0,i}$ is the outcome for person i when she does not receive the treatment.
- Known as potential outcomes.

Causal Identification

Ideal data

The *ideal* data for 10 people

```
#>      i trt  y1i  y0i effect_i
#> 1    1   1 5.01 2.56     2.45
#> 2    2   1 8.85 2.53     6.32
#> 3    3   1 6.31 2.67     3.64
#> 4    4   1 5.97 2.79     3.18
#> 5    5   1 7.61 4.34     3.27
#> 6    6   0 7.63 4.15     3.48
#> 7    7   0 4.75 0.56     4.19
#> 8    8   0 5.77 3.52     2.25
#> 9    9   0 7.47 4.49     2.98
#> 10 10   0 7.79 1.40     6.39
```

Calculate the causal effect of treatment.

$$\tau_i = y_{1,i} - y_{0,i}$$

for each individual i.

The mean of τ_i is the **average treatment effect** (ATE).

Thus, $\bar{\tau} = 3.82$

Notice the assignment of treatment does not influence ATE. Information on who was treated is irrelevant in this setting.

Fundamental Problem of Econometrics

Ideal comparison

$$\tau_i = y_{1,i} - y_{0,i}$$

Highlights the fundamental problem of econometrics, much like when a traveller assesses options down two separate roads.

The problem

- If we observe $y_{1,i}$, then we cannot observe $y_{0,i}$.
- If we observe $y_{0,i}$, then we cannot observe $y_{1,i}$.
- Can only observe what actually happened; cannot observe the **counterfactual**.

The traveller's alternative outcome is forever unknown to them.

Fundamental Problem of Econometrics

A dataset that we can observe for 10 people looks something like

```
#>      i trt  y1i  y0i
#> 1    1   1 5.01   NA
#> 2    2   1 8.85   NA
#> 3    3   1 6.31   NA
#> 4    4   1 5.97   NA
#> 5    5   1 7.61   NA
#> 6    6   0   NA 4.15
#> 7    7   0   NA 0.56
#> 8    8   0   NA 3.52
#> 9    9   0   NA 4.49
#> 10 10   0   NA 1.40
```

We can't observe $y_{1,i}$ and $y_{0,i}$.

But, we do observe

- $y_{1,i}$ for i in 1, 2, 3, 4, 5
- $y_{0,i}$ for i in 6, 7, 8, 9, 10

Q: How do we "fill in" the `NA`s and estimate $\overline{\tau}$?

Estimating Causal Effects

Notation: D_i is a binary indicator variable such that

- $D_i = 1$ if individual i is treated.
- $D_i = 0$ if individual i is not treated (*control* group).

Then, rephrasing the previous slide,

- We only observe $y_{1,i}$ when $D_i = 1$.
- We only observe $y_{0,i}$ when $D_i = 0$.

Q: How can we estimate $\bar{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Estimating Causal Effects

Q: How can we estimate $\bar{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

Idea: What if we compare the group of n peoples' means? *I.e.*,

$$\begin{aligned} &= Avg_n(y_i \mid D_i = 1) - Avg_n(y_i \mid D_i = 0) \\ &= Avg_n(y_{1i} \mid D_i = 1) - Avg_n(y_{0i} \mid D_i = 0) \end{aligned}$$

Q: When does a simple difference-in-means provide information on the **causal effect** of the treatment?

$Q_{2.0}$: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\bar{\tau}$?

Estimating Causal Effects

Assumption: Let $\tau_i = \tau$ for all i.

- The treatment effect is equal (constant) across all individuals i.

Note: We defined

$$\tau_i = \tau = y_{1,i} - y_{0,i}$$

which implies

$$y_{1,i} = y_{0,i} + \tau$$

Q: Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for τ?

Difference-in-means

$$= Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$$

$$= Avg(y_{1,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= Avg(\tau + y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \tau + Avg(y_{0,i} \mid D_i = 1) - Avg(y_{0,i} \mid D_i = 0)$$

$$= \text{Average causal effect} + \text{Selection bias}$$

Our proposed difference-in-means estimator gives us the sum of

1. τ, the **causal, average treatment effect** that we want.
2. **Selection bias:** How much treatment and control groups differ, on average.

Randomized Control Trials

Selection Bias

Problem: Existence of selection bias precludes *all else equal* comparisons.

- To make valid comparisons that yield causal effects, we need to shut down the bias term.

Potential solution: Conduct an experiment.

- How? **Random assignment of treatment**.
- Hence the name, ***randomized* control trial** (RCT).

Groups will need to be large enough to be comparable. As we saw with small sample extraction, the resulting estimated parameters may be far off from their population values. Following the LLN, as we increase n of both groups, our randomly assigned treatment estimate is **more likely** to be representative of the population mean.

Example: Effect of de-worming on attendance

Motivation: Intestinal worms are common among children in less-developed countries. The symptoms of these parasites can keep school-aged children at home, disrupting human capital accumulation.

Policy Question: Do school-based de-worming interventions provide a cost-effective way to increase school attendance?

Randomized Control Trials

Example: Effect of de-worming on attendance

Research Question: How much do de-worming interventions increase school attendance?

Q: Could we simply compare average attendance among children with and without access to de-worming medication?
A: If we're after the causal effect, probably not.

Q: Why not?
A: Selection bias: Families with access to de-worming medication probably have healthier children for other reasons, too (wealth, access to clean drinking water, *etc.*).
Can't make an *all else equal* comparison. Biased and/or spurious results.

Example: Effect of de-worming on attendance

Solution: Run an experiment.

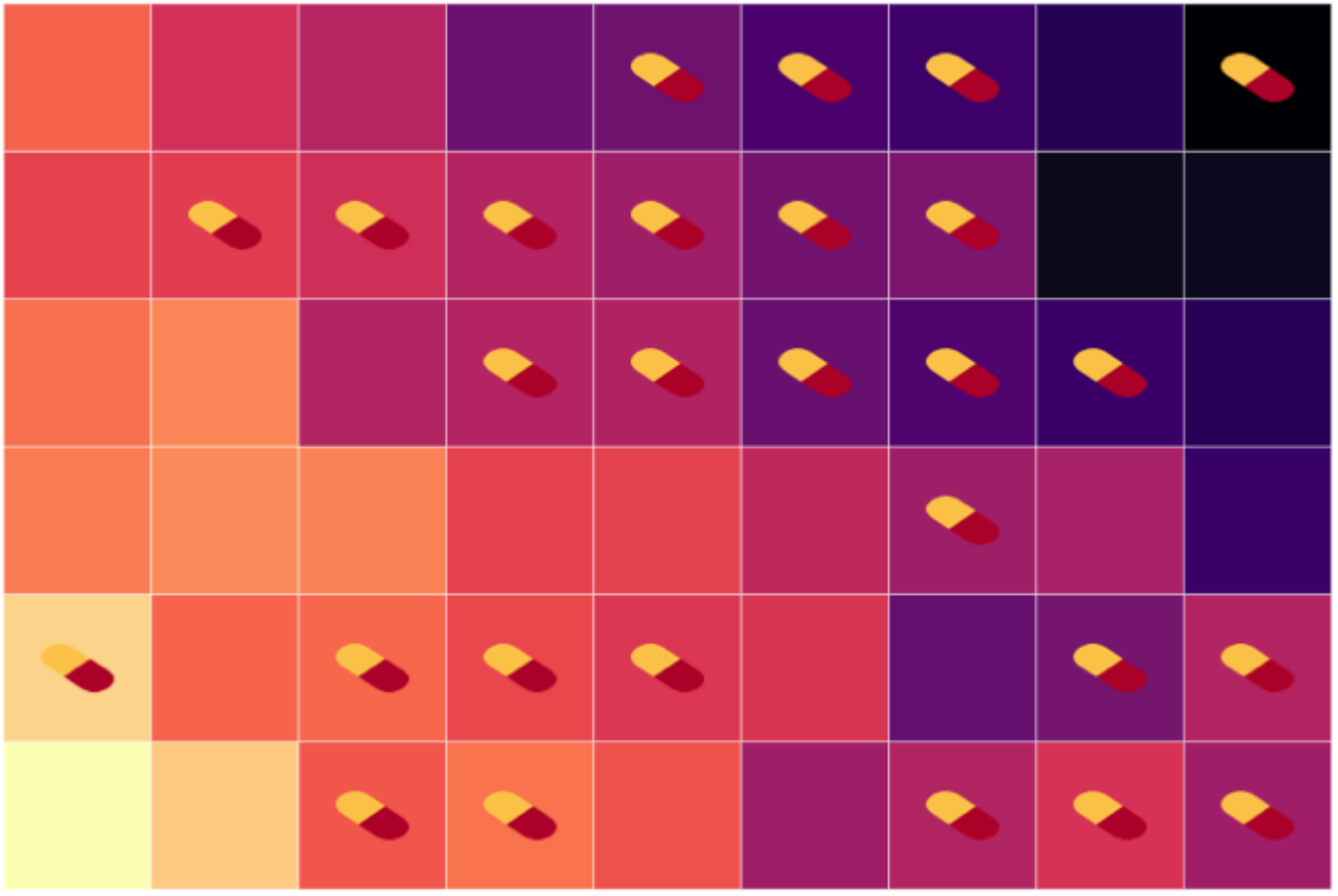
Imagine an RCT where we have two groups:

- **Treatment:** Villages where children get de-worming medication in school.
- **Control:** Villages where children don't get de-worming medication in school (status quo).

By randomizing villages into **treatment** or **control**, we will, on average, include all kinds of villages (poor *vs.* less poor, access to clean water *vs.* contaminated water, hospital *vs.* no hospital, *etc.*) in both groups.

All else equal!

54 villages of varying levels of development plus randomly assigned treatment



Randomized Control Trials

Example: Effect of de-worming on attendance

We can estimate the **causal effect** of de-worming on school attendance by comparing the average attendance rates in the treatment group (💊) with those in the control group (no 💊).

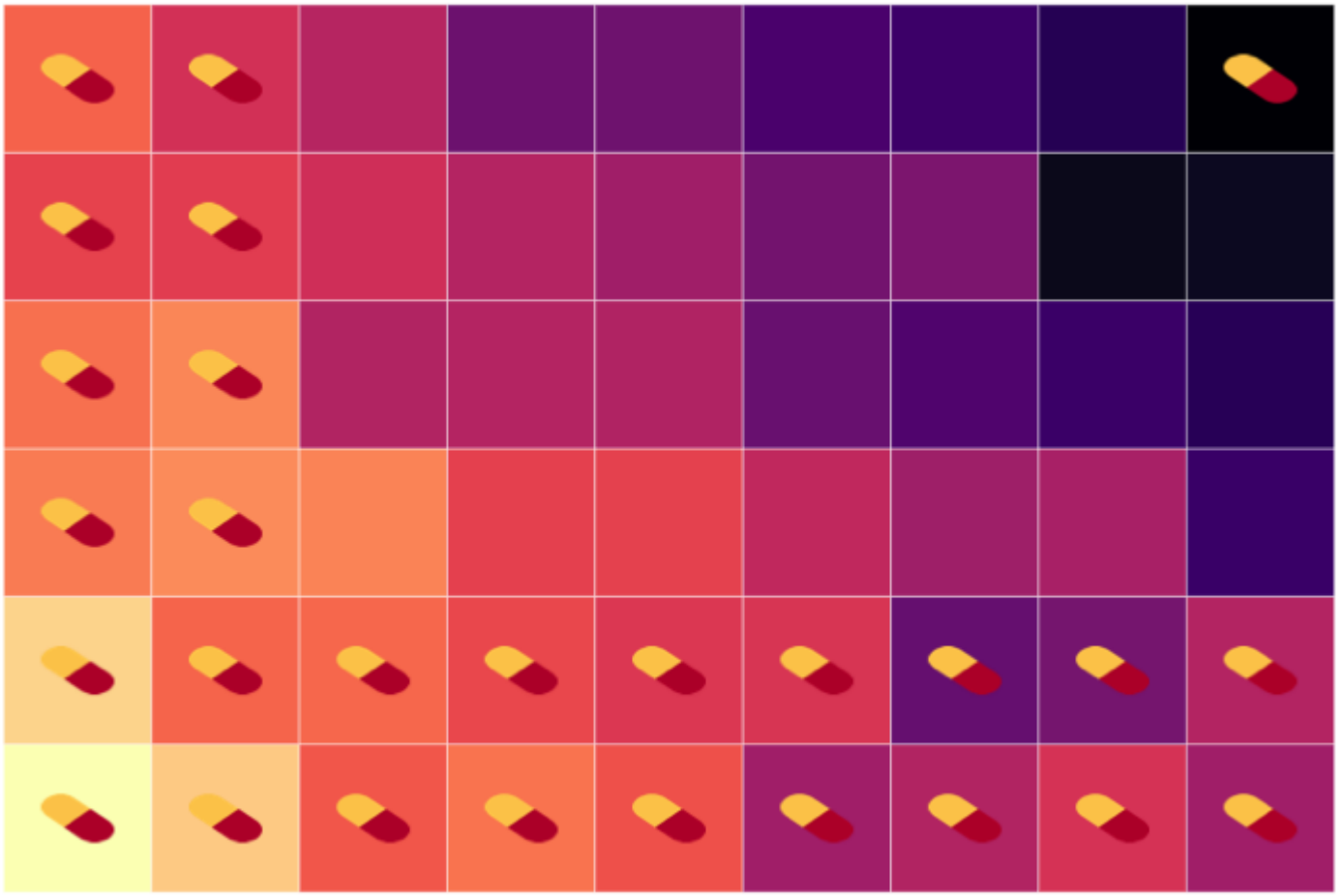
$$\overline{\text{Attendance}}_{\text{Treatment}} - \overline{\text{Attendance}}_{\text{Control}}$$

Alternatively, we can use the regression

$$\text{Attendance}_i = \beta_0 + \beta_1 \text{Treatment}_i + u_i \tag{1}$$

where Treatment_i is a binary variable (=1 if village i received the de-worming treatment). **Q:** Should trust the results of (1)? Why?
A: On average, **randomly assigning treatment should balance** treatment and control across the other dimensions that affect school attendance.

Randomization can go wrong!



Example: Returns to education

The optimal investment in education by students, parents, and legislators depends in part on the monetary *return to education*.

Thought experiment:

- Randomly select an individual.
- Give her an additional year of education.
- How much do her earnings increase?

The change in her earnings describes the **causal effect** of education on earnings.

Example: Returns to education

Q: Could we simply compare the earnings those with more education to those with less?
A: If we want to measure the causal effect, probably not.

1. People *choose* education based on their ability and other factors.
2. High-ability people tend to earn more *and* stay in school longer.
3. Education likely reduces experience (time out of the workforce).

Point (3) also illustrates the difficulty in learning about the effect of education while *holding all else constant*.

Many important variables have the same challenge: gender, race, income.

Example: Returns to education

Q: How can we estimate the returns to education?

Option 1: Run an **experiment**.

- Randomly assign education (might be difficult).
- Randomly encourage education (might work).
- Randomly assign programs that affect education (*e.g.*, mentoring).

Option 2: Look for a ***natural experiment*** (a policy or accident in society that arbitrarily increased education for one subset of people).

- Admissions cutoffs
- Lottery enrollment and/or capacity constraints

Summary

Takeaway

- The fundamental problem of econometrics is our inability to observe the outcome of a particular unit if assigned to treatment and control at once
- This introduces challenges in finding a way to construct a valid counterfactual using group means
- This quasi-experimental approach is valid if there is no selection bias or if we control for selection bias such that it is eliminated
- RCTs and natural experiments are one way to avoid selection bias.
- Since these events are often rare or infeasible, we will explore many other ways to approach this challenge