# Causality

## EC 421, Set 10

Connor Lennon
Spring 2021

# Prologue

# Schedule

## Last Time

Autocorrelation and nonstationarity

## Today

Causality

## Upcoming

**Assignment** Coming today; due early next week.

# Causality

# Causality

## Intro

Most tasks in econometrics boil down to one of two goals:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

1. **Prediction:** Accurately and dependably predict/forecast $y$ using on some set of explanatory variables—doesn't need to be $x_1$ through $x_k$. Focuses on $\hat{y}$. $\beta_j$ doesn't really matter.

2. **Causal estimation:**[†] Estimate the actual data-generating process—learning about the true, population model that explains how $y$ changes when we change $x_j$—focuses on $\beta_j$. Accuracy of $\hat{y}$ is not important.

For the rest of the term, we will focus on **causally estimating** $\beta_j$.

[†] Often called *causal identification*.

# Causality

## The challenges

As you saw in the data-analysis exercise, determining and estimating the true model can be pretty difficult—both practically and econometrically.

**Practical challenges**

- Which variables?
- Which functional form(s)?
- Do data exist? How much?
- Is the sample representative?

**Econometric challenges**

- Omitted-variable bias
- Reverse causality
- Measurement error
- How precise can/must we be?

Many of these challenges relate to **exogeneity**, *i.e.*, $\boldsymbol{E}[u_i|X] = 0$.
Causality requires us to **hold all else constant** (*ceterus paribus*).

# Causality

## It's complicated

Occasionally, ***causal*** relationships are simply/easily understood, *e.g.*,

- What caused the wildfire?
- How did this baby get here? 👶

Generally, ***causal*** relationships are complex and challenging to answer, *e.g.*,

- What causes some countries to grow and others to decline?
- Did pandemic aid payments cause crypto-coins to increase in price?
- Did lax regulation cause Texas's recent energy problems?
- How does the number of police officers affect crime?
- What is the effect of better air quality on test scores?
- Do longer/harsher prison sentences decrease crime?
- How did cannabis legalization affect mental health/opioid addiction?

# Causality

## Exogeneity

Another concept we need to think about is **exogeneity**.

When we start thinking about more complex systems of causality, we'll get into this more

but, for now, exogeneity is an assumption that some causal effect is unrelated to other causal factors.

We'll make this definition more concrete soon (and provide examples).

# Causality

## Correlation ≠ Causation

You've likely heard the saying

> Correlation is not causation. The saying is just pointing out that there are violations of exogeneity.

Although correlation is not causation, in an OLS model: **causation *requires* correlation**.

**New saying:**

> Correlation plus exogeneity is causation.

Let's work through a few examples.

# Causation

## Example: The causal effect of fertilizer[†][†]

Suppose we want to know the causal effect of fertilizer on corn yield.

**Q:** Could we simply regress yield on fertilizer?
**A:** Probably not (if we want the causal effect).

**Q:** Why not?
**A:** Omitted-variable bias: Farmers may apply less fertilizer in areas that are already worse on other dimensions that affect yield (soil, slope, water).
Violates *all else equal* (exogeneity). Biased and/or spurious results.

**Q:** So what *should* we do?
**A:** **Run an experiment!** 💩

[†] Many of the early statistical and econometric studies involved agricultural/ag-product field trials.
[†]Student's **t** actually comes from a statistician, William Gosset at guinness 🍺

# Causation

## Example: The causal effect of fertilizer

Randomized experiments help us maintain *all else equal* (exogeneity) so long as the randomization is un-associated with other causal factors.

We often call these experiments ***randomized control trials*** (RCTs).[†]

Imagine an RCT where we have two groups:

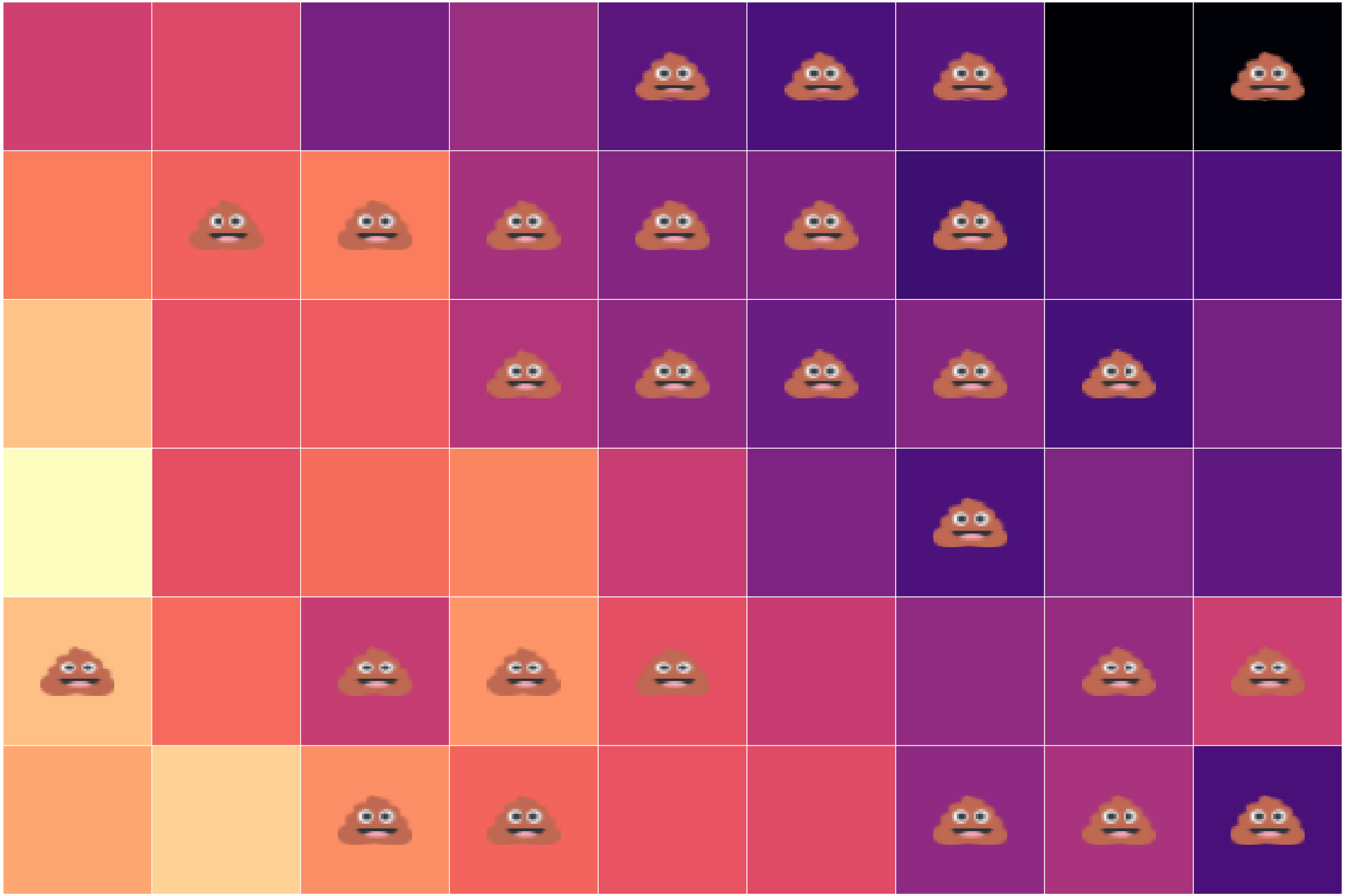
- **Treatment:** We apply fertilizer.
- **Control:** We do not apply fertilizer.

By randomizing plots of land into **treatment** or **control**, we will, on average, include all kinds of land (soil, slope, water, *etc.*) in both groups.

*All else equal*!

[†] Econometrics (and statistics) borrows this language from biostatistics and pharmaceutical trials.

# 54 equal-sized plots of varying quality plus randomly assigned treatment



# Causation

## Example: The causal effect of fertilizer

We can estimate the **causal effect** of fertilizer on crop yield by comparing the average yield in the treatment group (💩) with the control group (no 💩).

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

Alternatively, we can use the regression

$$\textbf{Yield}_i = \beta_0 + \beta_1 \textbf{Trt}_i + u_i \tag{1}$$

where $\textbf{Trt}_i$ is a binary variable (=1 if plot $i$ received the fertilizer treatment).

**Q:** Should we expect $(1)$ to satisfy exogeneity? Why?
**A:** On average, **randomly assigning treatment should balance** trt. and control across the other dimensions that affect yield (soil, slope, water).

# Causality

## Example: Returns to education

Labor economists, policy makers, parents, and students are all interested in the (monetary) *return to education.*

**Thought experiment:**

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

This change in earnings gives the **causal effect** of education on earnings.

# Causality

## Example: Returns to education

**Q:** Could we simply regress earnings on education?

**A:** Again, probably not if we want the true, causal effect.

1. People *choose* education based upon many factors, *e.g.*, ability.
2. Education likely reduces experience (time out of the workforce).
3. Education is ***endogenous*** (violates *exogeneity*).

The point (2) above also illustrates the difficulty in learning about educations while *holding all else constant*.

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

# Causality

## Example: Returns to education

**Q:** So how can we estimate the returns to education?

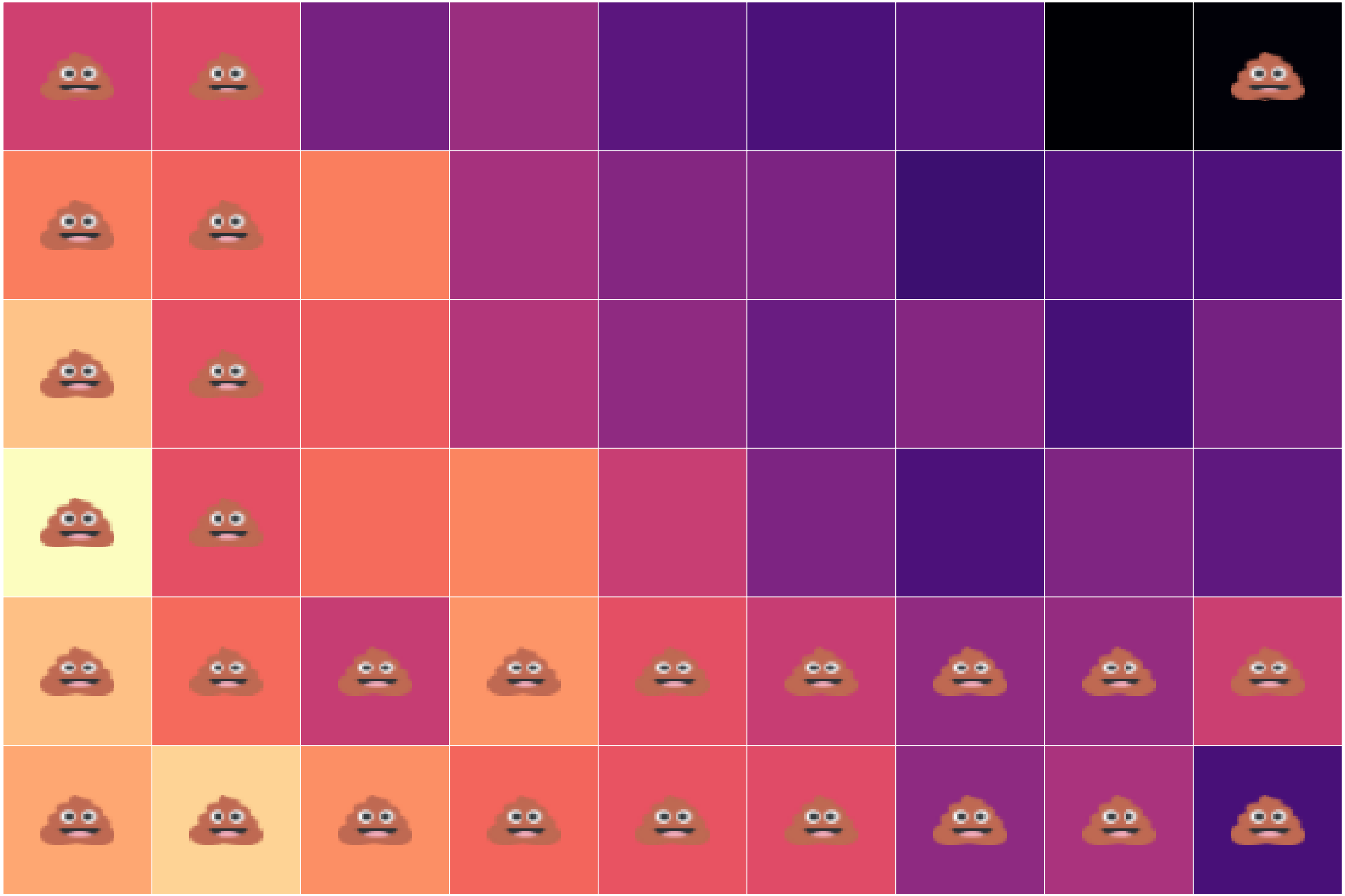
**Option 1:** Run an **experiment**.

- 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

# Unfortunate randomization



## The ideal experiment

The **ideal experiment** would be subtly different.

Rather than comparing units randomized as treatment vs. control, the ideal experiment would compare treatment and control **for the same, exact unit**.

$$y_{\text{Treatment},i} - y_{\text{Control},i}$$

which we will write (for simplicity) as

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

This *ideal experiment* is clearly infeasible[†], but it creates nice notation for causality (the Rubin causal model/Neyman potential outcomes framework).

[†] Without (1) God-like abilities and multiple universes or (2) a time machine.

# Causality

## The ideal experiment

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 trt.

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

for each individual $i$.

The mean of $\tau_i$ is the **average treatment effect** (ATE).

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

# Causality

## The ideal experiment

This model highlights the fundamental problem of causal inference.

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

**The challenge:**

If we observe $y_{1,i}$, then we cannot observe $y_{0,i}$.
If we observe $y_{0,i}$, then we cannot observe $y_{1,i}$.

## The ideal experiment

So a dataset that we actually observe for 6 people will look 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
- $y_{0,j}$ for $j$ in 6, 7, 8, 9, 10

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

# Causality

## Causally estimating the treatment effect

**Notation:** Let $D_i$ be 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 $\overline{\tau}$ using only $(y_{1,i}|D_i = 1)$ and $(y_{0,i}|D_i = 0)$?

# Causality

## Causally estimating the treatment effect

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

**Idea:** What if we compare the groups' means? *I.e.*,

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

**Q:** When does this simple difference in groups' 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 $\overline{\tau}$?

Time for math! 🎉

# Causality

## Causally estimating the treatment effect

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

This assumption says that 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$_{3.0}$:** Is $Avg(y_i \mid D_i = 1) - Avg(y_i \mid D_i = 0)$ a *good* estimator for $\tau$?

Difference in groups' means

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

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

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

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

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

So our proposed group-difference estimator give us the sum of

1. $\tau$, the **causal, average treatment effect** that we want
2. **Selection bias:** How much trt. and control groups differ (on average).

**Next time:** What if our causal system is more complex than that?

# Table of contents

## Admin

1. Schedule

## Causality

1. Introduction
2. The challenges
3. Examples
   - Fertilizer
   - Returns to education
4. *Real* experiments
5. The ideal experiment
6. Estimation
7. Derivation