Causal Inference: An Introduction

POLI 803 Research Methods in PS

Howard Liu

Fall 2025

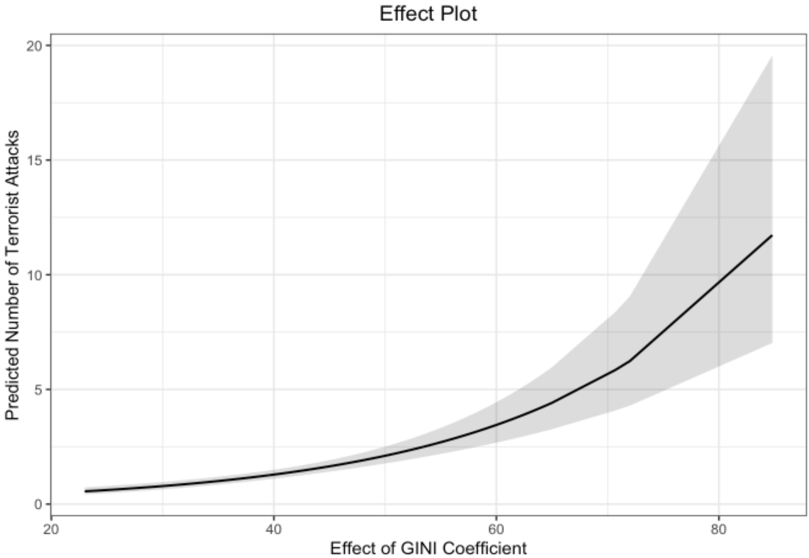
Roadmap

Foundational ideas

Some useful notations

Independence and Selection Bias

Statistical Inference (eg., Income Inequality)



We use the data we have to compute the average treatment effect

Economics Nobel Prize 2024

EKONOMIPRISET 2024
THE PRIZE IN ECONOMIC SCIENCES 2024

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES

Photo: Massachusetts Institute of Technology

Daron Acemoglu
Massachusetts Institute of Technology, USA

Photo: Massachusetts Institute of Technology

Simon Johnson
Massachusetts Institute of Technology, USA

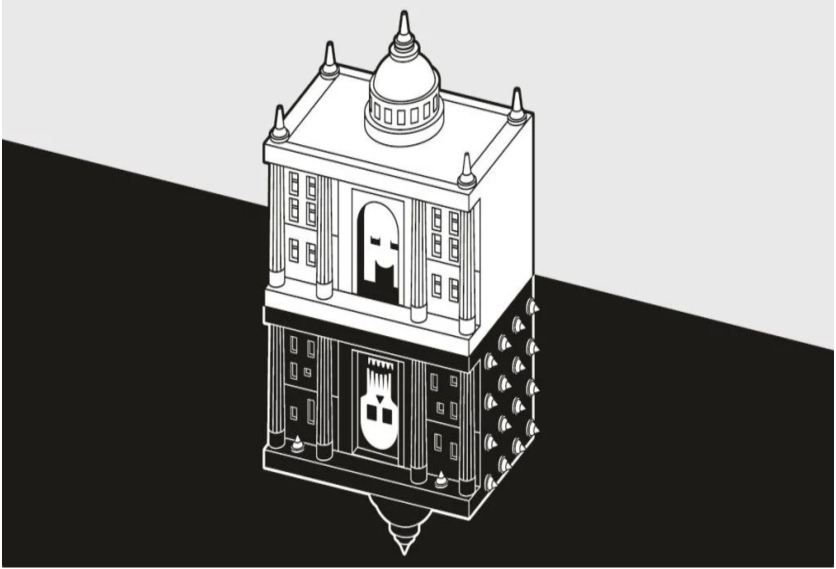
Photo: University of Chicago

James A. Robinson
University of Chicago, IL, USA

"för studier av hur institutioner formas och påverkar välstånd"

"for studies of how institutions are formed and affect prosperity"

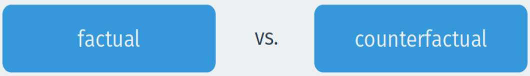
Economics Nobel Prize 2024 and Natural Experiments



Caveat: Before we dive in

- Causal inference is just one way of asking and answering research questions
- Remember: do not allow methods to drive your research question

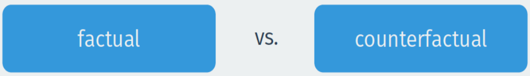
What is causal inference?



factual vs. counterfactual

- Does the minimum wage increase the unemployment rate?
  - → Unemployment rate went up after the minimum wage increased
  - → A causal question (a potential outcome/counterfactual framework):

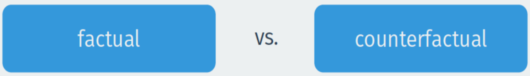
What is causal inference?



factual vs. counterfactual

- Does the minimum wage increase the unemployment rate?
  - → Unemployment rate went up after the minimum wage increased
  - → A causal question (a potential outcome/counterfactual framework): But would it have gone up if the minimum wage increase ”not” occurred?

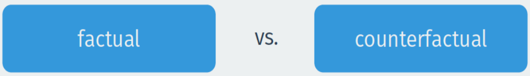
What is causal inference?



factual vs. counterfactual

- Does the minimum wage increase the unemployment rate?
  - → Unemployment rate went up after the minimum wage increased
  - → A causal question (a potential outcome/counterfactual framework): But would it have gone up if the minimum wage increase ”not” occurred?
- Does having girls affect a judge’s rulings in court?
  - → A judge with a daughter gave a pro-choice ruling
  - → A causal question (a potential outcome/counterfactual framework):

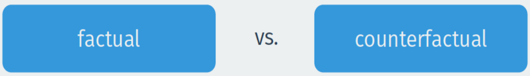
What is causal inference?



factual vs. counterfactual

- Does the minimum wage increase the unemployment rate?
  - → Unemployment rate went up after the minimum wage increased
  - → A causal question (a potential outcome/counterfactual framework): But would it have gone up if the minimum wage increase ”not” occurred?
- Does having girls affect a judge’s rulings in court?
  - → A judge with a daughter gave a pro-choice ruling
  - → A causal question (a potential outcome/counterfactual framework): Would they have done that if he had ”a son” instead?

What is causal inference?



factual vs. counterfactual

- Does the minimum wage increase the unemployment rate?
  - → Unemployment rate went up after the minimum wage increased
  - → A causal question (a potential outcome/counterfactual framework): But would it have gone up if the minimum wage increase ”not” occurred?
- Does having girls affect a judge’s rulings in court?
  - → A judge with a daughter gave a pro-choice ruling
  - → A causal question (a potential outcome/counterfactual framework): Would they have done that if he had ”a son” instead?
- Causal inference is the study of these types of causal questions by asking the **counterfactuals**
- The potential outcomes framework → build and discuss methods that estimate unbiased and meaningful **average** causal parameters as defined by that framework

What is not causal inference?

What is not causal inference?

- Correlations: parameters of the joint distribution of **observed data**
  - → Associations, regression coefficients, odds ratios, etc.
  - → Describes the world as it happened.
  - → No meaningful "directionality", just a joint distribution.
- Causal questions are about **unobserved data**: counterfactuals!

What is not causal inference?

Prediction

- Prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy (ML does great on this)
- Q: “Does this picture contain a human face or animal face? (Image AI)”, “How many wars will happen next year (forecasting)?”
- It’s not predictions of what ”effect” X will have on Y

What is not causal inference?

Prediction

- Prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy (ML does great on this)
- Q: “Does this picture contain a human face or animal face? (Image AI)”, “How many wars will happen next year (forecasting)?”
- It’s not predictions of what ”effect” X will have on Y

Causal inference

- Causal inference is also a type of prediction, but it’s a prediction of a *counterfactual* associated with a particular *choice taken*

What is not causal inference?

Prediction

- Prediction seeks to detect patterns in data and fit functional relationships between variables with a high degree of accuracy (ML does great on this)
- Q: “Does this picture contain a human face or animal face? (Image AI)”, “How many wars will happen next year (forecasting)?”
- It’s not predictions of what ”effect” X will have on Y

Causal inference

- Causal inference is also a type of prediction, but it’s a prediction of a *counterfactual* associated with a particular *choice taken*
- Causal inference takes that predicted (or imputed) counterfactual and constructs a causal effect that we hope tells us about a future in the event of a similar choice taken

#1: Correlation and causality are different concepts

- Causal question: “If a doctor puts a patient on a ventilator (D), will her covid symptoms (Y) improve?”
- Correlation question (correlation coefficient): Are ventilators correlated with improved covid symptoms?

$$\frac{Cov(D,Y)}{\sqrt{Var_D}\sqrt{Var_Y}}$$

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster ”cause” the sun to rise? Or did the sun cause the rooster to crow?

Modeling is Not the First Step

Most of us simply estimate models and cross our fingers that that coefficient is causal, but is it? When is it? Why is it? And which causal effect is it? And when is it reasonable to believe it?

We have to introduce concepts and notation first otherwise we will extend the correlation fallacy

Definition and Identification Come First

1. Turn the research question (“what is the causal effect of an advertising campaign on sales?”) into a specific aggregate causal parameter
2. Describe the narrow set of beliefs that make that parameter obtainable with data
3. Build a model that uses the data and the beliefs to estimate the causal parameter?

Most of us skip (1) and many skip (2) and go straight to (3).

Roadmap

Foundational ideas

Some useful notations

Independence and Selection Bias

Potential Outcomes Framework

- Econometrics traditionally modeled causality in terms of **realized outcomes** until recently (with some exceptions)

Potential Outcomes Framework

- Econometrics traditionally modeled causality in terms of **realized outcomes** until recently (with some exceptions)
- We need to make a distinction between now the idea of data ("realized outcomes") and **these hypothetical concepts ("potential outcomes")**

Potential outcomes notation

Let the treatment be a binary variable:

$$D_{i,t} = \begin{cases} 1 \text{ if placed on ventilator at time } t \\ 0 \text{ if not placed on ventilator at time } t \end{cases}$$

where i indexes an individual observation, such as a person

Potential outcomes notation

Potential outcomes:

$$Y^{j}_{i,t} = \begin{cases} 1 \text{ health if placed on ventilator at time } t \\ 0 \text{ health if not placed on ventilator at time } t \end{cases}$$

where j indexes a potential treatment status for the same i person at the same t point in time (pre or post treatment status)

Models vs Treatment Assignment

- Treatment assignment mechanism $(0, 1)$ drives the entire effort to identify causal effects as some make it easy and some make it potentially *impossible*

Models vs Treatment Assignment

- Treatment assignment mechanism (0, 1) drives the entire effort to identify causal effects as some make it easy and some make it potentially *impossible*
- Put another way, the same model can be unbiased and biased depending on the treatment assignment and be utterly detectable otherwise

Models vs Treatment Assignment

- Treatment assignment mechanism (0, 1) drives the entire effort to identify causal effects as some make it easy and some make it potentially *impossible*
- Put another way, the same model can be unbiased and biased depending on the treatment assignment and be utterly detectable otherwise
- Means modeling does not come first – it comes last

Important definitions

Definition 1: Individual treatment effect

The individual treatment effect, δ_i, associated with a ventilator is equal to $Y_i^1 - Y_i^0$.

Important definitions

Definition 2: Switching equation

An individual's realized health outcome, Y_i, is determined by treatment assignment, D_i which selects one of the potential outcomes:

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1-D_i)Y_i^0 \\
Y_i &= \begin{cases} Y_i^1 \text{ if } D_i = 1 \\ Y_i^0 \text{ if } D_i = 0 \end{cases} \\
\delta_i &= Y_i^1 - Y_i^0
\end{aligned}
$$

Missing data problem

Definition 3: Fundamental problem of causal inference

Since it is impossible to observe both Y_i^1 and Y_i^0 for the same individual, δ_i, is *unknowable*.

This is super important and why we take an average among aggregated individuals.

Missing data problem

- Causal inference = missing data problem
- Fundamental problem of causal inference is deep and impossible to overcome – not even with more data (you will always with more data be missing one of the potential outcomes)
- All of causal inference involves imputing missing counterfactuals and not all imputations are equal
- So what is the solution?

Missing data problem

- Causal inference = missing data problem
- Fundamental problem of causal inference is deep and impossible to overcome – not even with more data (you will always with more data be missing one of the potential outcomes)
- All of causal inference involves imputing missing counterfactuals and not all imputations are equal
- $\underline{\text{So what is the solution?}}$ → aggregating the treatment effect from individuals to the population, and then get an average

Average Treatment Effects

Definition 4: Average treatment effect (ATE)

The average treatment effect is the population average of all i individual treatment effects

$$
\begin{aligned}
E[\delta] &= E[Y^1 - Y^0] \\
&= E[Y^1] - E[Y^0]
\end{aligned}
$$

Aggregate parameters based on individual treatment effects are *summaries* of individual treatment effects

Cannot be calculated because Y_i^1 and Y_i^0 do not exist *for the same unit i* due to switching equation

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the ”treatment group” is equal to the average treatment effect conditional on being a treatment group member: (simply conditional probability)

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

Conditional Average Treatment Effects

Definition 5: Average Treatment Effect on the Treated (ATT)

The average treatment effect on the ”treatment group” is equal to the average treatment effect conditional on being a treatment group member: (simply conditional probability)

$$\begin{aligned} E[\delta|D=1] &= E[Y^1 - Y^0|D=1] \\ &= E[Y^1|D=1] - E[Y^0|D=1] \end{aligned}$$

- Note: remember to covid example (Y^1 = healthy, Y^0 = unhealthy)
- We compare effects before and after treatment in the treated group.
- Cannot be calculated for each individual because Y_i^1 and Y_i^0 do not exist **for the same unit i** due to switching equation.

Conditional Average Treatment Effects

Definition 6: Average Treatment Effect on the Untreated (ATU)

The average treatment effect on the ”untreated group” is equal to the average treatment effect conditional on being untreated:

$$\begin{aligned} E[\delta|D=0] &= E[Y^1 - Y^0|D=0] \\ &= E[Y^1|D=0] - E[Y^0|D=0] \end{aligned}$$

- We compare effects before and after treatment in the control group.
- Cannot be calculated because Y_i^1 and Y_i^0 do not exist **for the same unit i** due to switching equation.

Average Treatment Effects are Simple Summaries

- Theoretically, because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either

Average Treatment Effects are Simple Summaries

- Theoretically, because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Two major problems when moving from individual treatment effects to population treatment effects
  - → **Selection bias**: Individuals will ”choose” to take treatment based on the gain they expect from it → people select themselves into (or outside) treatment
  - → **Heterogenous treatment effect**: Every individual is different. The effect on the superman may not happen on me.

Average Treatment Effects are Simple Summaries

- Theoretically, because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Two major problems when moving from individual treatment effects to population treatment effects
  - → **Selection bias**: Individuals will ”choose” to take treatment based on the gain they expect from it → people select themselves into (or outside) treatment
  - → **Heterogenous treatment effect**: Every individual is different. The effect on the superman may not happen on me.
- So what do we do? Are we screwed?

Average Treatment Effects are Simple Summaries

- Theoretically, because aggregate causal parameters are *summaries* of individual treatment effects, each of which cannot be calculated, the aggregates cannot be calculated either
- Two major problems when moving from individual treatment effects to population treatment effects
  - → **Selection bias**: Individuals will ”choose” to take treatment based on the gain they expect from it → people select themselves into (or outside) treatment
  - → **Heterogenous treatment effect**: Every individual is different. The effect on the superman may not happen on me.
- So what do we do? Are we screwed?
- The magic: randomized sample! While we cannot **measure** average causal effects, we can **estimate** them when treatment is randomized in the sample we select

Simple Comparisons

Definition 7: Simple difference in mean outcomes (SDO)

A simple difference in mean outcomes (SDO) can be approximated by comparing the sample average outcome for the treatment group $(D = 1)$ with a comparison group $(D = 0)$

$$SDO = E[Y^1|D = 1] - E[Y^0|D = 0]$$

SDO is not a causal parameter because it's comparing Y^1 and Y^0 for different units, not the same units, so what is it measuring? What do we do with individual differences?

Estimate SDO with sample averages

$$
\begin{aligned}
\underbrace{E_{NT}[Y|D=1] - E_{NC}[Y|D=0]}_{\text{Estimate of SDO}} &= \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} \\
&+ \underbrace{E[Y^0|D=1] - E[Y^0|D=0]}_{\text{Selection bias}} \\
&+ \underbrace{(1-\pi)(ATT - ATU)}_{\text{Heterogenous treatment effect bias}}
\end{aligned}
$$

- The left-hand side can now be estimated because it's an average
- What is the right-hand side?

Estimate SDO with sample averages

$$
\begin{aligned}
\underbrace{E_{NT}[Y|D=1]-E_{NC}[Y|D=0]}_{\text{Estimate of SDO}} \quad = \quad & \underbrace{E[Y^1]-E[Y^0]}_{\text{Average Treatment Effect}} \\
& + \underbrace{E[Y^0|D=1]-E[Y^0|D=0]}_{\text{Selection bias}} \\
& + \underbrace{(1-\pi)(ATT-ATU)}_{\text{Heterogenous treatment effect bias}}
\end{aligned}
$$

- The left-hand side can now be estimated because it's an average
- What is the right-hand side? ATE + two biases
- Selection bias $\rightarrow$ issue with endogeneity
- Heterogenous treatment effect bias $\rightarrow$ issue with unit traits

Roadmap

Foundational ideas

Some useful notations

Independence and Selection Bias

Design vs model based approaches to selection bias

- Historically two ways that this selection bias was addressed: modeling it directly (Heckman, and others) and by design
    1. Design-based methods. Think of the randomized experiment. As we will see randomization will force selection bias to zero
    2. Model-based methods. Model the selection bias and then remove it mechanically
- Both have been highly influential, but constitute different approaches, and we largely focuses on the former not the latter

Selection bias and Design

- Source of the bias is caused by why some people get treated but others don't? Or some called the "treatment assignment *mechanism*"

Treatment assignment mechanisms

- Two extreme examples of a treatment assignment mechanism:
    1. randomization (i.e., taking the medicine because a coin flip told you to take it) versus
    2. sorting on one or both potential outcome (i.e., taking the medicine because it'll help you) which we call the "Perfect Doctor" but which Heckman and others call a "Roy model"

Treatment assignment mechanisms

- Two extreme examples of a treatment assignment mechanism:
    1. randomization (i.e., taking the medicine because a coin flip told you to take it) versus
    2. sorting on one or both potential outcome (i.e., taking the medicine because it'll help you) which we call the "Perfect Doctor" but which Heckman and others call a "Roy model"
- Bias comes from how treatment is assigned and that mechanism dictates the direction we have to take

Three forms of selection bias (because we are human)

- In causal inference, selection bias is caused by different mean potential outcomes by treatment status, of which there are three possibilities:
  1. Selection on Y^0: You chose the treatment because of something will happen if you didn't (good health if having vaccine protection)
  2. Selection on Y^1: You chose the treatment because of something will happen if you do (will be protected)
  3. Selection on gains, δ: You chose the treatment because the net benefits were positive

Three forms of selection bias (because we are human)

- In causal inference, selection bias is caused by different mean potential outcomes by treatment status, of which there are three possibilities:
    1. Selection on Y^0: You chose the treatment because of something will happen if you didn't (good health if having vaccine protection)
    2. Selection on Y^1: You chose the treatment because of something will happen if you do (will be protected)
    3. Selection on gains, δ: You chose the treatment because the net benefits were positive
- All three cause biased estimates of the ATE, though the degree to which it fully distorts the estimates depends on those different reasons for sorting into treatment

Summarizing the goals of causal inference

Our goal in causal inference is to estimate aggregate causal parameters with data using treatment assignment mechanisms that plausibly eliminate selection bias

Summarizing the goals of causal inference

Our goal in causal inference is to estimate aggregate causal parameters with data using treatment assignment mechanisms that plausibly eliminate selection bias

Depending on the treatment assignment mechanism, certain procedures are allowed and others are prohibited

Let's look what happens in an RCT *and why* this addresses selection bias term $E[Y^0|D=1]$ and $E[Y^0|D=0]$ to see why Fisher (1925) recommended it

Independence (D and Y)

Independence assumption

Treatment is assigned to a population independent of that population's potential outcomes

$$(Y^0, Y^1) \perp\!\!\!\perp D$$

This is random or quasi-random assignment and ensures mean potential outcomes for the treatment group and control group are the same. Also ensures other variables are distributed the same for a large sample.

$$\begin{aligned} E[Y^0|D=1] &= E[Y^0|D=0] \\ E[Y^1|D=1] &= E[Y^1|D=0] \end{aligned}$$

Random Assignment Solves the Selection Problem

$$\begin{aligned}\underbrace{E_N[y_i|d_i=1]-E_N[y_i|d_i=0]}_{\text{SDO}} &= \underbrace{E[Y^1]-E[Y^0]}_{\text{Average Treatment Effect}}\\ &+\underbrace{E[Y^0|D=1]-E[Y^0|D=0]}_{\text{Selection bias}}\\ &+\underbrace{(1-\pi)(ATT-ATU)}_{\text{Heterogenous treatment effect bias}}\end{aligned}$$

- If treatment is independent of potential outcomes (meaning no individuals sorted themselves into treatment), then swap out equations and **selection bias** zeroes out!

$$E[Y^0|D=1]-E[Y^0|D=0] = 0$$

Random Assignment Solves the Heterogenous Treatment Effects

- How does randomization affect heterogeneity treatment effects bias from the third line? Rewrite definitions for ATT and ATU:

$$\text{ATT} = E[Y^1|D=1] - E[Y^0|D=1]$$
$$\text{ATU} = E[Y^1|D=0] - E[Y^0|D=0]$$

- Rewrite the third row bias after $1-\pi$:

$$\begin{aligned} ATT - ATU &= \mathbf{E[Y^1 \mid D{=}1]} - E[Y^0|D=1] \\ &\quad - \mathbf{E[Y^1 \mid D{=}0]} + E[Y^0|D=0] \\ &= 0 \end{aligned}$$

Identification with Full Independence

$$\underbrace{E_N[Y_i|D_i=1] - E_N[Y_i|D_i=0]}_{\text{Estimate of SDO}} = \underbrace{E[Y^1] - E[Y^0]}_{\text{Average Treatment Effect}} + \underbrace{0}_{\text{Selection bias}} + \underbrace{0}_{\text{Heterogenous treatment effect bias}}$$

SDO is unbiased estimate of ATE with randomized treatment assignment because it sets selection bias to zero and $ATT = ATU$.

Interference when aggregating units

- While treatment effects are defined at individual level, aggregate parameters combine units

Interference when aggregating units

- While treatment effects are defined at individual level, aggregate parameters combine units
- → This therefore means that for the aggregate parameters to be stable, one unit's treatment choice cannot "**interfere**" with another unit's potential outcomes
- Huge headaches, even in the RCT
- Violations are an active area of scholarship and important for social networks, peer effects and various platforms (e.g., Twitter) → spatial, temporal, network interdependence

Interference when aggregating units

- While treatment effects are defined at individual level, aggregate parameters combine units
- → This therefore means that for the aggregate parameters to be stable, one unit's treatment choice cannot "**interfere**" with another unit's potential outcomes
- Huge headaches, even in the RCT
- Violations are an active area of scholarship and important for social networks, peer effects and various platforms (e.g., Twitter) → spatial, temporal, network interdependence
- Scholars work on theoreizing interference (i.e., diffusion effects) rather than just controlling or ignoring them

SUTVA

- SUTVA stands for “stable unit-treatment value assumption”
    1. **S**: ***s****table*
    2. **U**: across all ***u****nits*, or the population
    3. **TV**: ***t****reatment-value* (“treatment effect”, “causal effect”)
    4. **A**: ***a****ssumption*

SUTVA assumptions and violations

What are some possible violations you can imagine?

SUTVA: No spillovers to other units

- What if we impose a treatment at one neighborhood but not a contiguous one?
- Treatment may spill over causing $Y = Y^1$ even for the control units because of spillovers from treatment group

SUTVA: No spillovers to other units

- What if we impose a treatment at one neighborhood but not a contiguous one?
- Treatment may spill over causing $Y = Y^1$ even for the control units because of spillovers from treatment group
- Can be mitigated with careful delineation of treatment and control units so that interference is impossible, may even require aggregation (e.g., classroom becomes the unit, not students)

SUTVA: No Hidden Variation in Treatment

- SUTVA requires each unit receive the same treatment dosage; this is what it means by “stable” (i.e., notice that the super scripts contain either 0 or 1, not 0.55, 0.27)
- If we are estimating the effect of aspirin on headaches, we assume treatment is 200mg per person in the treatment
- Easy to imagine violations if hospital quality, staffing or even the vents themselves vary across treatment group
- Be careful what we are and are not defining as *the treatment*; you may have to think of it as multiple arms

SUTVA: Scale can affect stability of treatment effects

Easier to imagine this with a different example.

- Let’s say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?

SUTVA: Scale can affect stability of treatment effects

Easier to imagine this with a different example.

- Let’s say we estimate a causal effect of early childhood intervention in Texas
- Now President Biden wants to roll it out for the whole United States – will it have the same effect as we found?
- Scaling up a policy can be challenging to predict if there are rising costs of production
- What if expansion requires hiring lower quality teachers just to make classes?
- That’s a general equilibrium effect; we only estimated a partial equilibrium effect (external versus internal validity)