

# Econometrics of Policy Evaluation: Causality

Cristian Huse

# Outline

- ① The Policy Question: Correlation vs. Causation
- ② The Causal Framework: Potential Outcomes & Causal Diagrams
- ③ The Core Problem: Selection Bias
- ④ The Ideal Solution: Randomization
- ⑤ Real-World Challenges: Assumptions & Threats to Validity

- Policy questions are causal in nature
  - Does one more year of education cause higher income? (\*)
  - Does a change in bankruptcy law lower interest rates?
  - Does calorie posting in restaurants reduce calorie consumption?
  - Does school decentralization improve school quality?
- Problem is, the statistics you have learned in college does not address this... Why?
- Example (\*):
  - Hypothesized causal effect we want to measure: *Education → Income*
  - But what about *Inate Ability*? It is a non-causal source of correlation between *Education* and *Income*
  - Graphically: *Education ← Inate Ability → Income*, or a “back-door” path
  - **Challenge of policy evaluation:** isolate the causal pathway of interest from the **confounding** back-door pathways
- **Note:** A Causal Diagram (DAG) is a graphical tool for structure how we think about causality (a “language of causality”)

# Standard Statistical Analysis

- Tools
  - Likelihood, OLS and other estimation techniques
- Aim
  - To infer parameters of a distribution (or DGP = Data Generating Process) from samples drawn of that distribution
- With the help of such parameters, one can
  - Infer association among variables
  - Estimate the likelihood of past and future events
  - Update the likelihood of events given new evidence/measurement
- But it describes associations within a **stable DGP** – for this to work well, **experimental conditions must remain the same**
  - Now recall our policy questions...
    - If I make a child go to school longer, will s/he earn more money?
    - If I change the bankruptcy law, will interest rates decrease?
  - **The conditions change!**

- For causal questions, we need to infer aspects of the DGP
- We need to be able to deduce:
  - ① the likelihood of events under **static conditions** (as in Standard Statistical Analysis)
  - ② the dynamics of events under **changing conditions**
- “**dynamics of events under changing conditions**” includes:
  - ① Predicting the effects of interventions
  - ② Predicting the effects of spontaneous changes
  - ③ Identifying causes of reported events

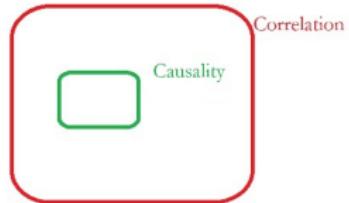
- Standard statistical analysis/probability theory:
  - “Causality” is not in its vocabulary
  - i.e. only allows us to say that two events are mutually correlated, or dependent ( $\neq$ causal)
- This is not enough for policy makers
  - Policy makers want a relation of cause and effect
  - They look at rationales for policy decisions
    - i.e. if we do X, then will we get Y?
  - To do so, we first need a vocabulary for causality...
- **Example (\*):**
  - The simple correlation between education and income reflects the sum of two paths:
  - The direct causal path ( $Education \rightarrow Income$ ); and
  - The confounding back-door path  
 $(Education \leftarrow Innate\ Ability \rightarrow Income)$
  - Causal analysis is the work of separating these two.

- First randomized experiment was in psychometrics (Peirce and Jastrow, 1885)
- Philip Wright discovered the instrumental variables estimator in 1928
- Trygve Haavelmo won the Nobel Prize in 1989 for his work on “simultaneous equations” in which he showed that *regression cannot identify supply and demand simultaneously from a series of price and quantity bundles because a regression of intersections between supply and demand won’t identify whether the supply curve or the demand curve had shifted*
- Roland Fisher and Jerzy Neyman proposed the **Potential Outcomes Framework**, and its powerful notation, in the 1920s and Donald Rubin revived it for the social sciences in the 1970s

- John Stuart Mill: “*If a person eats of a particular dish, and dies in consequence, that is, would not have died if he had not eaten it, people would be apt to say that eating of that dish was the source of his death.*”
- Roland Fisher: “*If we say, ‘This boy has grown tall because he has been well fed,’ ... we are suggesting that he might quite probably have been worse fed, and that in this case he would have been shorter.*”
- Stock and Watson: “*A causal effect is defined to be the effect of a given action or treatment, as measured in an ideal, randomized controlled experiment. In such an experiment, the only systematic reason for differences in outcomes between the treatment and control groups is the treatment itself.*” → **Hints at the most credible solution to answer a causality question**
- Haavelmo (1944): “*What makes a piece of mathematical economics not only math but also economics is, I believe, this: When we set up a system of theoretical relationships and use economic names for the otherwise purely theoretical variables involves, we have in mind some actual experiment, or some design of an experiment, which we could at least imagine arranging, in order to measure those quantities in real economic life that we think might obey the laws imposed on their theoretical namesakes.*”



“Correlation is not Causality” vs “Causality is Correlation”:



- A lot of the language in the experimental and quasi-experimental literature is borrowed from the medical literature (e.g., “treatment”, “control”)
- Simple example introducing **potential outcomes** notation and the selection problem: *“Do hospitals make people healthier?”*
- National Health Interview Survey (NHIS) 2005
  - Health status measured from 1 (poor health) to 5 (excellent health)

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7,774	3.21	0.014
No hospital	90,049	3.93	0.003

# Notation

- Think of hospitalization as a “treatment” and health status as an “outcome”
- Let the treatment (hospitalization) be a binary variable denoted

$$D_i = \begin{cases} 1 & \text{if hospitalized} \\ 0 & \text{if not hospitalized} \end{cases}$$

where  $i$  indexes an individual observation, such as a person

- Observed outcomes (health status) is  $Y_i$  but individual is either hospitalized or not
- **Causal question:**  $D \rightarrow Y$ 
  - i.e., “does hospitalization ( $D_i$ ) cause health ( $Y_i$ )?”
- Contrast with “is hospitalization correlated with health?”
- What’s the difference between the two questions? Is that an trivial or meaningful distinction in your opinion?

# Notation

- Correlation as a merely statistical concept:  $\frac{1}{N} \frac{\text{Cov}(D, Y)}{\sqrt{\text{Var}(D)} \sqrt{\text{Var}(Y)}}$ 
  - i.e., variables move together
  - but so do drownings in swimming pools and #Nicholas Cage movie appearances within a year, plus many other bizarre variables
- Causation has a deeper meaning...
  - Policy  $D$  has an effect on educational, environmental, health outcome  $Y$
  - (go back to quotes in early slides)

- In Mexico, a conditional cash transfer (CCT) programme called Progresa/Oportunidades led to the following:

Group	Sample Size	Mean Pre-Program HH Income
Received CCT	10,000	USD 150/month
Did Not Receive	40,000	USD 400/month

- Did the Progresa program make people poorer? Or was it **targeted** at the poorest households?
- This creates the exact same selection problem as sick people selecting into hospitals.

- Observed variables:

- Treatment –  $D_i$  – is observed as **either** 0 or 1 for each  $i$  unit.
- Actual outcomes are observed for each unit

- Unobserved variables:

- We can't observe any of the  $i$  unit's **counterfactual** (potential) outcomes
- For each individual unit,  $i$ , there exist two potential outcomes associated with our binary treatment, hospitalization:

$$\text{Potential outcome} = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases}$$

- $Y_i(0)$  is the health status of an individual had she not gone to the hospital – **regardless of whether she did in fact go to the hospital**
- $Y_i(1)$  is the health status of an individual had she gone to the hospital – **regardless of whether she did in fact go to the hospital**

**Definition 1.** The **individual treatment effect**,  $\delta_i$ , equals  $Y_i(1) - Y_i(0)$

**Definition 2.** The **average treatment effect (ATE)** is the population average of all  $i$  individual treatment effects

$$\begin{aligned} E[\delta_i] &= E[Y_i(1) - Y_i(0)] \\ &= E[Y_i(1)] - E[Y_i(0)] \end{aligned}$$

**Definition 3.** An individual's observed , say, health outcome,  $Y$ , is determined by treatment assignment,  $D_i$ , and corresponding potential outcomes via the **switching equation**:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

$$Y_i = \begin{cases} Y_i(1) & \text{if } D_i = 1 \\ Y_i(0) & \text{if } D_i = 0 \end{cases}$$

**Definition 4. Fundamental Problem of Causal Inference.** It is **impossible** to observe both  $Y_i^1$  and  $Y_i^0$  for the same individual. So, individual causal effects –  $\delta_i$  – are impossible to know.

- **In words:** The FPCI is a [data problem](#)
  - We are missing half of the potential outcomes data for every single unit in our sample.
  - This is why causal inference is fundamentally different from descriptive statistics: it requires making assumptions to fill in this missing data.
- How can one bridge this “gap of knowledge”?
  - Some studies have used identical twins...
  - But typically one will need data + assumptions
- **Example.** If authorities implement a lockdown during a pandemic, one can only observe what happened under the lockdown policy – i.e., what actually happened – not what would have happened in the absence of a lockdown
  - Consequence: if the policy is successful (few cases, few deaths) some people might argue that the pandemic was a hoax to start with

The best way to go through this example is to use pencil and paper and take some notes

- Consider the following two situations:
  - ① Jack is in the hospital ( $D_{Jack} = 1$ ) and his health is a 2 ( $Y_{Jack} = 2$ ).
  - ② Jill is not in the hospital ( $D_{Jill} = 0$ ) and her health is a 4 ( $Y_{Jill} = 4$ )
- According to Definition 3 (switching equation), we know the following:
  - ① Jack's observed health outcome,  $Y_{Jack} = 2$  equals his potential health outcome under treatment,  $Y_{Jack}^1 = 2$
  - ② Jill's observed health outcome,  $Y_{Jill} = 4$  equals her potential health outcome under "control",  $Y_{Jill}^0 = 4$

- According to Definitions 1 and 4, the following is also true:
  - We do not know Jack's potential health outcome under control,  $Y_{Jack}(0)$  and therefore we do not know the causal effect of hospitalization on Jack's health since  $\delta_{Jack} = Y_{Jack}(1) - Y_{Jack}(0)$
  - We do not know Jill's potential health outcome under treatment,  $Y_{Jill}(1)$  and therefore we do not know the causal effect of hospitalization on Jill's health since  $\delta_{Jill} = Y_{Jill}(1) - Y_{Jill}(0)$
  - For the very same reasons that we don't know the individual treatment effect, we do not know the average causal effect,  $E[\delta]$ , because we are missing  $Y_{Jill}(1)$  and  $Y_{Jack}(0)$  – the **missing counterfactuals** (Definition 4) – both of which are needed to calculate ATE
- We cannot calculate the ATE because we are missing individual treatment effects, and we are missing individual treatment effects because we are missing counterfactuals for each unit  $i$

**Definition 5.** The **average treatment effect on the treated (ATT)** is equal to the average treatment effect conditional on being a treatment group member:

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

**Definition 6.** The **average treatment effect on the untreated (ATU)** is equal to the average treatment effect conditional on being untreated:

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

- The **FPCI** (Fundamental Problem of Causal Inference) is that since we cannot observe both  $Y_{Jack}(1)$  and  $Y_{Jack}(0)$  (or any  $i$  for that matter) at the same moment in time, then we cannot calculate the causal effect of hospitalization on health outcomes
- We therefore have to rely on *observable health outcomes*,  $Y$ , since potential health outcomes are not completely available to us,  $Y(1), Y(0)$
- But if we subtract Jill's observed health outcome ( $Y_{Jill} = 4$ ) from Jack's observed health outcome, ( $Y_{Jack} = 2$ ), we get:

$$(Y_{Jack}|D_{Jack} = 1) - (Y_{Jill}|D_{Jill} = 0) = 2 - 4 = -2$$

which implies that hospitalization *caused* health to go from good to poor

- How does this relate to the FPCI?

- Back to hospitalization example...
  - Do hospitalizations make people sick? Or do sick people go to hospitals?
  - This is called the *selection problem* – without the full potential outcomes, the simple difference between treatment and control group won't tell us the causal effect of hospitalization on health outcomes
  - Same problem carries over to the simple difference in means,  
 $E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$
- So what are we actually measuring if we *naively* compare average health status for the hospitalized with that of the non-hospitalized?

**Definition 7.** A **simple difference in mean outcomes (SDO)** is the difference between the population average outcome for the treatment and control groups, and can be approximated by the sample averages:

$$E[Y(1)|D = 1] - E[Y(0)|D = 0] = E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0]$$

in large samples. It is sometimes also called the **naive average treatment effect**

- Notice that:
  - All individuals in the population contribute twice to ATE, whereas a sampled individual is used only once to estimate SDO by contributing to either  $E_N[y_i|d_i = 1]$  or  $E_N[y_i|d_i = 0]$ .
  - Statistical models, such as SDO, are valuable insofar as they can provide unbiased and/or consistent estimates of the parameter of interest (i.e., ATE).

- In practice, the SDO is what one would get from:
  - A simple t-test between treated and non-treated sub-samples;
  - A bivariate regression of the outcome on the treatment dummy.
- What is the difference between SDO and ATE?

$$\begin{array}{ccc} SDO & \leq & ATE \\ E[Y(1)|D=1] - E[Y(0)|D=0] & \geq & E[Y(1)] - E[Y(0)] \end{array}$$

- The LHS term is the *estimator* of the *parameter* on the RHS, and estimators can be biased.
- Our goal:
  - Learn about a feature of the world (the parameter ATE) using data to compute an estimator (the SDO).
- Our central question:
  - Under what conditions is this estimator unbiased for the parameter we care about?

- Back to the hospitalization example, our challenge is that sickness affects both hospitalization and health status:



where  $U$  is an unobserved variable for underlying health status.

- Using the SDO amounts to ignoring the role of the  $U$  variable in the causal diagram above (“open back-door path”)
- Intuitively, what we want is a way to isolate the rectangle above to quantify the horizontal arrow effect (causal effect)

$Hospital (D) \rightarrow Health (Y)$

- The simple difference in mean outcomes (SDO) can be decomposed into three terms (ignoring sample average notation) (proof in Appendix):

$$\begin{aligned} E[Y(1)|D=1] - E[Y(0)|D=0] &= \text{ATE} \\ &\quad + E[Y(0)|D=1] - E[Y(0)|D=0] \\ &\quad + (1-\pi)(\text{ATT} - \text{ATU}) \end{aligned} \quad (1)$$

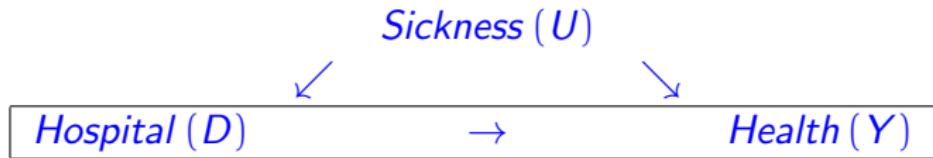
where  $\pi$  is the proportion of the population receiving treatment.

- How do we interpret this?

- $\text{ATE} = E[Y(1) - Y(0)]$ , is the parameter of interest
- $E[Y(1)|D=1] - E[Y(0)|D=0]$  is the **selection bias** term, the difference in potential health outcome for the treatment and control group had neither received any hospitalization (i.e., how did Jack and Jill's health status compare to one another had Jack *not* gone to hospital? Was he sicker than her *already*?)
- $(1-\pi)(\text{ATT} - \text{ATU})$  shows that part of the bias is due to heterogeneous treatment effects weighted by the share of the population in the control group – recall outcome is a distribution and first term is the average (i.e., different “amounts of sickness”)

- We have established that the simple difference in means is a biased estimate of the causal effect due to selection.
- The next section asks: how can we devise a strategy to eliminate this bias?

- **The problem:** How to remove the “confounding back-door path” below:



- **The solution:** Remove the ↘ arrow from confounder to treatment
- **How:** Using random assignment – a method where treatment status  $D$  is determined by a coin flip (or a computer), and not by any pre-existing characteristic of the individual, like  $U$ .
- **Consequence:**

$$E[Y(1)|D = 1] - E[Y(0)|D = 0] = ATE$$

- Let us make it formal...

**Definition: Independence assumption.** Treatment is independent of potential outcomes.

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

- **In words:** Random assignment (into treatment) means that the treatment has been assigned to units independent of their potential outcomes. Thus, mean potential outcomes for the treatment group and control group are the same **for a given state of the world**

$$E[Y(1)|D = 1] = E[Y(1)|D = 0]$$

$$E[Y(0)|D = 1] = E[Y(0)|D = 0]$$

- **Intuition:** Use a lottery to decide which individual gets treated vs. non-treated

- **Claim: Randomization solves the selection problem**
  - i.e., random assignment of  $D_i$  (treatment) makes treatment  $D_i$  **independent** of potential outcomes,  $Y_i^1$  and/or  $Y_i^0$
- What does “independence” mean exactly? If treatment is independent of potential outcomes, then the mean potential outcomes for the treatment group and control group are the same *for a given state of the world*:

$$E[Y_i(1)|D_i = 1] = E[Y(1)|D_i = 0]$$

$$E[Y_i(0)|D_i = 1] = E[Y(0)|D_i = 0]$$

- Notice that the selection bias from the second line of the decomposition of SDO was:

$$E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0]$$

- If treatment is independent of potential outcomes, then we can swap out the above equations and the selection bias will disappear:

$$\begin{aligned} E[Y_i(0)|D_i = 1] - E[Y_i(0)|D_i = 0] &= E[Y_i(0)|D_i = 0] - E[Y_i(0)|D_i = 0] \\ &= 0 \end{aligned}$$

- Randomization dealt with selection bias, but what about the heterogeneity treatment effects bias term,  $(1 - \pi)(ATT - ATU)$ ?
- **Claim: Randomization solves the heterogeneity problem**
- Rewrite definitions for ATT and ATU:

$$ATT = E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1]$$

$$ATU = E[Y_i(1)|D_i = 0] - E[Y_i(0)|D_i = 0]$$

- And rewrite the third line term after  $(1 - \pi)$ :

$$\begin{aligned} ATT - ATU &= E[Y_i(1)|D_i = 1] - E[Y_i(0)|D_i = 1] \\ &\quad - E[Y_i(1)|D_i = 0] + E[Y_i(0)|D_i = 0] \end{aligned}$$

- Use the independence assumption to make the heterogeneity vanish (homework!)
- **Conclusion:** If treatment is independent of potential outcomes, then:

$$\begin{aligned} E_N[y_i|d_i = 1] - E_N[y_i|d_i = 0] &= E[Y(1)] - E[Y(0)] \\ SDO &= ATE \end{aligned}$$

# What independence does not mean

- Notice – independent treatment assignment means that

$$E[Y(1)|D_i = 1] = E[Y(1)|D_i = 0]$$

and the equivalent for  $Y(0)$

- But that does not in any way imply that there is no causal effect. Independence does not imply, in other words, that  $E[Y(1)|D = 1]$  is equal to  $E[Y(0)|D = 0]$ .
- Independence only implies that the average values for a given potential outcome (i.e.,  $Y(1)$  or  $Y(0)$ ) are the same for the groups who did receive the treatment and those who did not

- The potential outcomes model presupposes a set of bundled assumptions called SUTVA – a jargon acronym short for “stable unit-treatment value assumption”
  - ① **S**: is *stable*
  - ② **U**: across all *units*, or the population
  - ③ **TV**: that the *treatment-value* (“treatment effect”, “causal effect”)
  - ④ **A**: SUTVA is an *assumption*
- This is an often overlooked assumption in causal estimation. In the simplest terms, SUTVA means that average causal effects are parameters that assume
  - ① Homogenous dosage; and
  - ② Potential outcomes (and any function of them) that are invariant to how many other people receive the treatment

# SUTVA: Homogeneous dose

- SUTVA constrains what the treatment can be
- Individuals are receiving the same treatment – i.e., the “dose” of the treatment to each member of the treatment group is the same. That’s the “stable unit” part
- If we are estimating the effect of hospitalization on health status, we assume everyone is getting the same dose of the hospitalization treatment
  - Easy to imagine violations if hospital quality varies, though, across individuals. But, that just means we have to be careful what we are and are not defining as *the treatment*
- If we are estimating the effect of a teacher training programme and some teachers receive intensive coaching while others get a simple brochure, the “treatment” is not stable, and we are estimating a confusing average of different effects.

## SUTVA: No externalities

- What if hospitalizing Jack (hospitalized,  $D = 1$ ) is actually about vaccinating Jack from small pox?
  - If Jack is vaccinated for small pox, then Jill's potential health status (without vaccination) may be higher than when he isn't vaccinated
  - i.e.,  $Y_{Jill}(0)$ , may vary with what Jack does *regardless of whether Jill herself receives treatment*
- Consider de-worming programmes (analogous for vaccines)
  - A famous study found that treating some children for intestinal worms in a school reduced the overall prevalence of worms, i.e., treatment benefitted children in the control group
  - This is a positive spillover and a clear violation of the “no interference between units” aspect of SUTVA.
  - This biases the simple SDO estimate of the treatment effect downwards.
- In sum, if there are no externalities from treatment, then  $\delta_i$  is stable for each  $i$  unit regardless of whether someone else receives the treatment too.

# SUTVA: Partial equilibrium only

- “**No general equilibrium effects**”
  - e.g., the effects of a program don’t spill over to other markets
- If a large-scale job training programme provides thousands of people with new skills (e.g., welding), it could increase the supply of welders so much that it drives down the market wage for all welders, even in different industries.
  - This “general equilibrium” effect impacts both the treated group and the control group, violating SUTVA and complicating the interpretation of the estimated effect.

## Example: Krueger (1999)

- Krueger (1999) econometrically re-analyzes a randomized experiment to determine the **causal effect of class size on student achievement**
- The project is Tennessee Student/Teacher Achievement Ratio (STAR) run in the 1980s
- 11,600 students and their teachers were **randomly** assigned to one of the following three groups:
  - ① Small classes of 13-17 students
  - ② Regular classes of 22-25 students
  - ③ Regular classes of 22-25 students with a full-time teacher's aide
- After the assignment, the design called for students to remain in the same class type for four years
- Randomization occurred within schools
- **With randomization one could simply calculate SDO – Why?**

- In the STAR experiment, randomization gives us the independence assumption. This means the selection bias is zero, and the ATE is identified by the Simple Difference in Outcomes
- Now that we have identified the parameter we want, we need to estimate it from our sample data:
  - Can use a simple t-test or, equivalent and more flexible, a regression.
- Estimating TEs using regression can be done as follows:
  - Assume that TEs are constant – i.e.,  $Y_i(1) - Y_i(0) = \delta \forall i$
  - Substitute into a rearranged switching equation (Definition 2):

$$\begin{aligned}
 Y_i &= D_i Y_i(1) + (1 - D_i) Y_i(0) \\
 Y_i &= Y_i(0) + (Y_i(1) - Y_i(0)) D_i \\
 Y_i &= Y_i(0) + \delta D_i \\
 Y_i &= E[Y_i(0)] + \delta D_i + (Y_i(0) - E[Y_i(0)]) \\
 Y_i &= \alpha + \delta D_i + \eta_i
 \end{aligned}$$

where  $\eta_i = (Y_i(0) - E[Y_i(0)])$  is the random part of  $Y_i(0)$

- Thus, can use a regression equation to estimate the causal effect of  $D$  on  $Y$

- The conditional expectation,  $E[Y_i|D_i]$ , with treatment status switched on and off gives:

$$E[Y_i|D_i = 1] = \alpha + \delta + E[\eta_i|D_i = 1]$$

$$E[Y_i|D_i = 0] = \alpha + E[\eta_i|D_i = 0]$$

- Subtracting the latter from the former, we get:

$$\underbrace{E[Y_i|D_i = 1] - E[Y_i|D_i = 0]}_{SDO} = \underbrace{\delta}_{\text{Treatment Effect}} + \underbrace{E[\eta_i|D_i = 1] - E[\eta_i|D_i = 0]}_{\text{Selection bias}}$$

- We can estimate  $SDO$  using least squares but there are other options as well
  - In the STAR experiment,  $D_i$ , equalled one if the student was enrolled in a small class and had been **randomly** assigned
  - Recall that randomization implies that treatment is independent of potential outcomes, and therefore the selection bias vanishes

# Why Include Control Variables?

- To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the above equation, we might estimate

$$Y_i = \alpha + \delta D_i + X'_i \gamma + \eta_i$$

- There are two main reasons for including additional controls in the regression models:
  - ① **Conditional random assignment.** Sometimes randomization is done **conditional** on some observable. (here that's the school)
  - ② **Additional controls increase precision.** Although control variables  $X_i$  are uncorrelated with  $D_i$ , they may have substantial **explanatory power** for  $Y_i$   
⇒ Including controls thus reduces residual variance and therefore lowers the standard errors of the regression estimates

# Why Include Control Variables?

- To evaluate experimental data, one may want to add additional controls in the multivariate regression model. So, instead of estimating the above equation, we might estimate

$$Y_i = \alpha + \delta D_i + X'_i \gamma + \eta_i$$

- There are two main reasons for including additional controls in the regression models:
  - ① **Conditional random assignment.** Sometimes randomization is done **conditional** on some observable. (here that's the school)
  - ② **Additional controls increase precision.** Although control variables  $X_i$  are uncorrelated with  $D_i$ , they may have substantial **explanatory power** for  $Y_i$   
⇒ Including controls thus reduces residual variance and therefore lowers the standard errors of the regression estimates

# Why Include Control Variables? Using Causal Diagrams

- **Conditional random assignment**

- In the STAR experiment, randomization was done *within schools*. This means *School* is a common cause of both class size assignment  $D$  and student achievement  $Y$ , leading to a back-door path:

$$D \leftarrow \text{School} \rightarrow Y$$

- By including school fixed effects in the regression, we are “controlling for” *School*, which visually means we are “blocking” this back-door path, satisfying the independence assumption conditional on school.

- **Additional controls increase precision**

- Assume randomization is perfect, so there are no back-door paths.
- A variable like *Prior Test Score* is a very strong predictor of the outcome  $Y$ , but is uncorrelated with  $D$  due to randomization. The DAG would be  $D [x] \text{Prior Score} \rightarrow Y$ .
- The error term  $\eta_i$  in a simple regression of outcome on treatment contains all unobserved factors affecting  $Y$ , including *Prior Test Score*.

# Main Results

- Recall

$$Y_i = \alpha + \delta D_i + X'_i \gamma + \eta_i$$

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
A. Kindergarten				
Small class	4.82 (2.19)	5.37 (1.26)	5.36 (1.21)	5.37 (1.19)
Regular/aide class	.12 (2.23)	.29 (1.13)	.53 (1.09)	.31 (1.07)
White/Asian (1 = yes)	—	—	8.35 (1.35)	8.44 (1.36)
Girl (1 = yes)	—	—	4.48 (.63)	4.39 (.63)
Free lunch (1 = yes)	—	—	-13.15 (.77)	-13.07 (.77)
White teacher	—	—	—	-.57 (2.10)
Teacher experience	—	—	—	.26 (.10)
Master's degree	—	—	—	-.51 (1.06)
School fixed effects	No .01	Yes .25	Yes .31	Yes .31
$R^2$				

Explanatory variable	OLS: actual class size			
	(1)	(2)	(3)	(4)
B. First grade				
Small class	8.57 (1.97)	8.43 (1.21)	7.91 (1.17)	7.40 (1.18)
Regular/aide class	3.44 (2.05)	2.22 (1.00)	2.23 (0.98)	1.78 (0.98)
White/Asian (1 = yes)	—	—	6.97 (1.18)	6.97 (1.19)
Girl (1 = yes)	—	—	3.80 (.56)	3.85 (.56)
Free lunch (1 = yes)	—	—	-13.49 (.87)	-13.61 (.87)
White teacher	—	—	—	-.43 (1.96)
Male teacher	—	—	—	11.82 (3.33)
Teacher experience	—	—	—	.05 (0.06)
Master's degree	—	—	—	.48 (1.07)
School fixed effects	No .02	Yes .24	Yes .30	Yes .30
$R^2$				

- Conditional random assignment: *School* fixed-effects
- Additional controls increase precision:
  - Note how std errors decrease and  $R^2$  increases from (1) to (2)
  - Little change of point estimates for “Small class”, i.e. be it for Kindergarten or First grade

# Empirical Challenges: Threats to Validity

- Brief comments on real-life challenges faced by Krueger...
- **Attrition**
  - i.e., people leaving the experiment
  - If attrition is non-random (e.g., weakest students in large classes/good students from large classes drop out/move to private schools), it breaks the independence assumption and re-introduces selection bias.
- **Heterogeneous treatment effects bias**
  - i.e., occurs if treatment effects differ across the population (i.e.,  $ATT \neq ATU$ )
  - Potential problem: If people selecting to take part of the randomized trial have different returns compared to the population average, then the experiment will only identify a localized average treatment effect (LATE) for the sub-population participants in the experiment

# Empirical Challenges II

- Imperfect Compliance/Switching classrooms

B. First grade to second grade

First grade	Second grade			
	Small	Regular	Reg/aide	All
Small	1435	23	24	1482
Regular	152	1498	202	1852
Aide	40	115	1560	1715
All	1627	1636	1786	5049

- i.e., non-zero off-diagonals in transition matrix
- e.g., students with stronger expected academic potential were more likely to move into small classes
- Potential problem: if students switched between TG and CG, then comparing them yields biased/inconsistent estimates
- Potential solution: IV estimation, where the initial random assignment is the “instrument”

- Supply Side Changes

- If programs are scaled up, the supply side implementing the treatment may be different
- In the trial phase, the supply side may be more motivated than during the large scale roll-out of a program

# Empirical Challenges III

- **Hawthorne effects**

- People behave differently if they are being observed in an experiment. Similar to “placebo effects” in that this is a false positive result
- If they operate differently on treatment and control groups, then they may introduce biases
- If people from the control group behave differently, these effects are sometimes called “John Henry” effects
- These are psychological spillovers where the control group’s behavior is affected by their awareness of the experiment, violating the “no interference” assumption.

- **Substitution bias**

- Control group members may seek substitutes for treatment
- This would bias the estimated treatment effects **downward** (Why?)
- Can also occur if the experiment frees up resources that can now be concentrated on the control group

## Take-aways

- Impact evaluations establish the extent to which a program – and that program alone – caused a change in an outcome
- The counterfactual is what would have happened – what the outcome ( $Y$ ) would have been for a program participant – in the absence of the program ( $P$ , or  $D$ )
- Since we cannot directly observe the counterfactual, we must estimate it
- Without a control (comparison) group that yields an accurate estimate of the counterfactual, the true impact of a program cannot be established
- In sum, causal inference requires a clear theoretical model of the world (a “Theory of Change” or Causal Diagram) to guide the choice of an identification strategy.

# Take-aways

- A valid control group...
  - ① has the same characteristics, on average, as the treatment group in the absence of the program;
  - ② remains unaffected by the program; and
  - ③ would react to the program in the same way as the treatment group, if given the program
- When the control group doesn't accurately estimate the true counterfactual, the estimated impact of the program will be biased
- **Selection bias** occurs when the reasons for which an individual participates in a program are correlated with outcomes
  - Ensuring that the estimated impact is free of selection bias is one of the major objectives and challenges for any impact evaluation
- **Randomization** as the gold standard when it comes to program evaluation. However...
  - It is not always feasible
  - It relies on a number of assumptions that need to be critically evaluated

# References

- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition. Washington, DC: Inter-American Development Bank and World Bank
  - chapter 3
- Gertler et al (2016). Impact Evaluation in Practice, 2nd. Edition, Technical Companion (Version 1.0). Washington, DC: Inter-American Development Bank and World Bank.
  - p. 2-5
- Huntington-Klein, N. (2022). The Effect: An Introduction to Research Design and Causality. Routledge
- [MW] Morgan and Winship (2014). Counterfactuals and Causal Inference.

## Examples of “Fake” Counterfactuals

- Think about the causal diagrams of the following two cases
- Before-after comparisons
  - Pre = no fertilizer vs Post = fertilizer
  - Causal effect: impact of fertilizer use on crop yields
  - Assume drought happens in Post period, so output halves, but is not observed by economist
  - False conclusion would be that fertilizer decreased output (!)
  - Conclusion: Unless can account for *every other factor* affecting output, cannot calculate the true impact of the program by using a before-after comparison
- Self-selection into (or out of) treatment
  - Treatment: vocational training for unemployed, free entry
  - Two years ahead, compare incomes of those who chose to enroll with those who chose not to enroll in training
  - Impact calculation says that incomes of treated twice as high as those who chose not to be treated
  - Problem: groups likely very different, e.g. motivation, ability

# Examples of “Fake” Counterfactuals

- Before-after comparisons
  - The confounder is *Time* (e.g., the drought happened in the post-period). The back-door path is  $Fertilizer \leftarrow Time \rightarrow Yields$ .
- Self-selection into (or out of) treatment
  - The confounder is unobserved *Motivation*. The back-door path is  $Training \leftarrow Motivation \rightarrow Income$ .

Some terms will be treated in more detail in later topics...

- **Average Treatment Effect.** The average treatment effect across the population.
- **Average Treatment on the Treated.** The average treatment effect among those who actually received the treatment in your study.
- **Average Treatment on the Untreated.** The average treatment effect among those who did not actually receive the treatment in your study.
- **Conditional Average Treatment Effect.** The average treatment effect among those with certain values of certain variables (for example, the average treatment effect among women).
- **Heterogeneous Treatment Effect.** A treatment effect that differs from individual to individual.

- **Intent-to-Treat.** The average treatment effect of assigning treatment, in a context where not everyone who is assigned to receive treatment receives it (and vice versa) →Estimated from randomized experiments with non-compliance.
- **Local Average Treatment Effect.** A weighted average treatment effect where the weights are based on how much more treatment an individual would get if assigned to treatment than if they weren't assigned to treatment →Estimated by Instrumental Variables.
- **Marginal Treatment Effect.** The treatment effect of the next individual that would be treated if treatment were expanded.
- **Weighted Average Treatment Effect.** A treatment effect average where each individual's treatment effect is weighted differently.
- **Variance-Weighted Average Treatment Effect.** A treatment effect average where each individual's treatment effect is weighted based on how much variation there is in their treatment variable, after closing back doors.

This appendix provides the formal algebraic derivation (Morgan and Winship (2014) , p. 26):) of the relationship between the SDO, the ATE, and bias from both selection and heterogeneous treatment effects. The core intuition for the selection bias component was illustrated in the main text using Causal Diagrams.

**Step 1: ATE is equal to sum of conditional average expectations by LIE**

$$\begin{aligned}\text{ATE} &= E[Y(1)] - E[Y(0)] \\ &= \{\pi E[Y(1)|D = 1] + (1 - \pi)E[Y(1)|D = 0]\} \\ &\quad - \{\pi E[Y(0)|D = 1] + (1 - \pi)E[Y(0)|D = 0]\}\end{aligned}$$

Use simplified notation:

$$\begin{aligned}E[Y(1)|D = 1] &= a \\ E[Y(1)|D = 0] &= b \\ E[Y(0)|D = 1] &= c \\ E[Y(0)|D = 0] &= d \\ \text{ATE} &= e\end{aligned}$$

and rewrite ATE as  $e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$

## Step 2: Move SDO terms to LHS

$$e = \{\pi a + (1 - \pi)b\} - \{\pi c + (1 - \pi)d\}$$

$$e = \pi a + b - \pi b - \pi c - d + \pi d$$

$$e = \pi a + b - \pi b - \pi c - d + \pi d + (\mathbf{a} - \mathbf{a}) + (\mathbf{c} - \mathbf{c}) + (\mathbf{d} - \mathbf{d})$$

$$= e - \pi a - b + \pi b + \pi c + d - \pi d - \mathbf{a} + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d} + \mathbf{d}$$

$$\mathbf{a} - \mathbf{d} = e - \pi a - b + \pi b + \pi c + d - \pi d + \mathbf{a} - \mathbf{c} + \mathbf{c} - \mathbf{d}$$

$$\mathbf{a} - \mathbf{d} = e + (\mathbf{c} - \mathbf{d}) + \mathbf{a} - \pi a - b + \pi b - \mathbf{c} + \pi c + d - \pi d$$

$$\mathbf{a} - \mathbf{d} = e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)a - (1 - \pi)b + (1 - \pi)d - (1 - \pi)c$$

$$\mathbf{a} - \mathbf{d} = e + (\mathbf{c} - \mathbf{d}) + (1 - \pi)(a - c) - (1 - \pi)(b - d)$$

### Step 3: Substitute conditional means

$$\begin{aligned}
 E[Y(1)|D = 1] - E[Y(0)|D = 0] &= \text{ATE} \\
 &\quad + (E[Y(0)|D = 1] - E[Y(0)|D = 0]) \\
 &\quad + (1 - \pi)(\{E[Y(1)|D = 1] - E[Y(0)|D = 1]\}) \\
 &\quad - (1 - \pi)\{E[Y(1)|D = 0] - E[Y(0)|D = 0]\}) \\
 E[Y(1)|D = 1] - E[Y(0)|D = 0] &= \text{ATE} \\
 &\quad + (E[Y(0)|D = 1] - E[Y(0)|D = 0]) \\
 &\quad + (1 - \pi)(ATT - ATU)
 \end{aligned}$$

## Step 4: Decomposition of difference in means

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

where  $E_N[y_i|d_i = 1] \rightarrow E[Y(1)|D = 1]$ ,  $E_N[y_i|d_i = 0] \rightarrow E[Y(0)|D = 0]$  and  $(1 - \pi)$  is the share of the population in the control group.