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* $\rightarrow$ *Income*
  - But what about *Inate Ability*? It is a non-causal source of correlation between *Education* and *Income*
  - Graphically: *Education* $\leftarrow$ *Inate Ability* $\rightarrow$ *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!

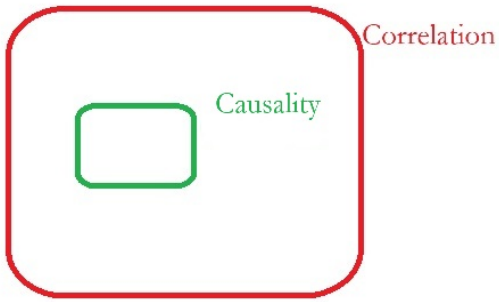
- For causal questions, we need to infer aspects of the DGP
- We need to be able to deduce:
  1. the likelihood of events under static conditions (as in Standard Statistical Analysis)
  2. the dynamics of events under changing conditions
- "dynamics of events under changing conditions" includes:
  1. 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 (≠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* → *Income*); and
  - The confounding back-door path (*Education* ← *Innate Ability* → *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:** $D \to Y$
  - 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: $\frac{1}{N} \frac{Cov(D,Y)}{\sqrt{Var(D)}\sqrt{Var(Y)}}$
  - 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:

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

  - $Y_i(0)$ 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**, δ_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**:

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

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 – δ_i – are impossible to know.

- **In words:** More than just a data problem, the FPCI is a logical 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:
  1. 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 $\delta_{Jack} = Y_{Jack}(1) - Y_{Jack}(0)$
  2. 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 $\delta_{Jill} = Y_{Jill}(1) - Y_{Jill}(0)$
  3. 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**

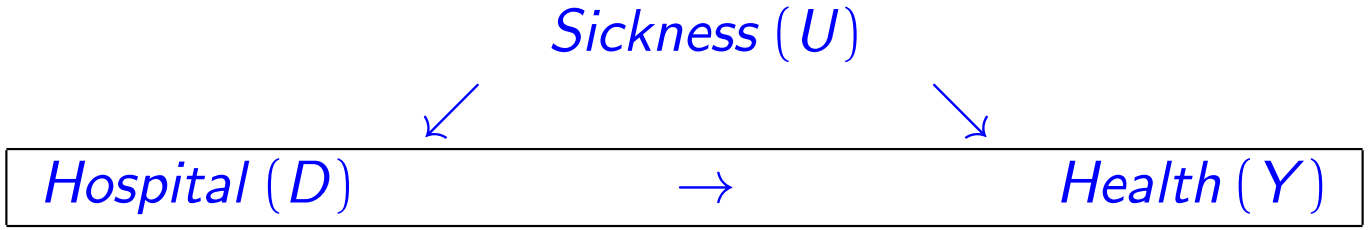
- Notice that:
  - All individuals in the population contribute twice to ATE, whereas a sampled individual is used only once to estimate SDO by contributing to either $E_N[y_i|d_i=1]$ or $E_N[y_i|d_i=0]$.
  - Statistical models, such as SDO, are valuable insofar as they can provide unbiased and/or consistent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

$$\begin{aligned} SDO &\underset{>}{\overset{\leq}{}} ATE \\ E[Y(1)|D=1] - E[Y(0)|D=0] &\underset{>}{\overset{\leq}{}} E[Y(1)] - E[Y(0)] \end{aligned}$$

  - 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)

$$\textit{Hospital}\,(D) \rightarrow \textit{Health}\,(Y)$$

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

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

  where π is the proportion of the population receiving treatment.

- How do we interpret this?
  $ATE = E[Y(1) - Y(0)]$, is the parameter of interest, the true **causal effect** we want to isolate. It answers the question: *"For a person chosen at random from the population, what is the average effect of the treatment?"*

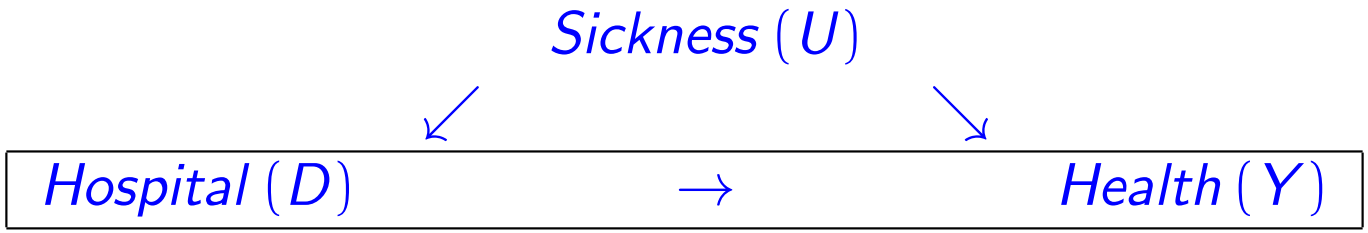
- How do we interpret this? (cont'd)

$E[Y(0)|D=1] - E[Y(0)|D=0]$ is the **selection bias** term, which captures pre-existing differences between the groups. It asks the crucial counterfactual question: *"Suppose no one had received the treatment. Would the treated and untreated groups have had different outcomes anyway?"*
In the hospital example, this is the formal way of asking: *"Were the people who went to the hospital already sicker than those who stayed home?"*
If this term is not zero, the simple comparison is biased.

$(1-\pi)(ATT - ATU)$ arises if the treatment has a different impact on different types of people. It captures the difference between the effect on the treated (ATT) and the effect on the untreated (ATU). It asks: *"Is the effect of hospitalization different for the kind of people who choose to go to the hospital versus the kind who don't?"*
If the effect is not uniform across the population, this also contributes to the overall bias.

- 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*:

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

- Under the independence assumption, one can use the last equation to rewrite the selection bias term as follows: :

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

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

which yields zero.

- 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 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

A Foundational Assumption: SUTVA

- Randomization is our "gold standard" for eliminating selection bias.
- However, the entire Potential Outcomes framework – and thus our interpretation of an RCT – relies on a crucial, often unstated, assumption about how treatments function in the world.
- This is the Stable Unit Treatment Value Assumption (SUTVA):
  1. **S**: is ***s****table*
  2. **U**: across all ***u****nits*, or the population
  3. **TV**: that the ***t****reatment-value* ("treatment effect", "causal effect")
  4. **A**: SUTVA is an ***a****ssumption*

Defining SUTVA

- SUTVA is a compound assumption with two critical components:
  1. **No Interference:** The potential outcomes for any individual depend only on their own treatment status. They are not affected by who else receives the treatment.
  2. **No Hidden Variations of Treatment:** The treatment assigned is identical for everyone who receives it. There is only one version of the treatment.
- **Why it matters:**
  - Without SUTVA, we cannot simply write $Y_i(1)$. We would need a more complex notation like $Y_i(D_i, D_{-i})$, where an individual's outcome depends on their own treatment (D_i) and the vector of treatments for everyone else (D_{-i}).
  - i.e., SUTVA simplifies the causal question to a manageable form.

Violations of SUTVA: No Hidden Variations

- **Example 1 (Teacher Training):**
  - A teacher training program is evaluated. Some teachers receive intensive, one-on-one coaching, while others just get a brochure.
  - If both are coded as "treated", we are averaging the effects of two different treatments, violating the "no hidden variations" assumption.
- **Example 2 (Hospital Quality):**
  - In the hospitalization example, if some patients go to a state-of-the-art research hospital and others to an under-resourced rural clinic, the "treatment" is not stable.

Violations of SUTVA: No Interference & Spillovers

- **Positive Spillovers:**
  - Treating some children for worms reduces the overall parasite load in a school, which also benefits children in the control group.
  - This spillover causes us to underestimate the true program effect, because the control group is partially treated.
- **Negative Spillovers (Displacement):**
  - A job training program helps participants find jobs. However, some of these jobs might have otherwise gone to individuals in the control group.
  - This "displacement" harms the control group, causing us to overestimate the net benefit of the program.
- **General Equilibrium Effects:**
  - If a large-scale agricultural program successfully increases crop yields, it could lower food prices for everyone, including the control group.
  - This market-level spillover complicates the measurement of the direct effect on farmers.

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:

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

- Subtracting the latter from the former, we get:

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

- 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. **Blocking "Backdoors":** Even in an RCT, randomization might be done within certain groups (e.g., within each school). This creates a backdoor path: $\textit{Treatment} \leftarrow \textit{School} \rightarrow \textit{Outcome}$. Including school fixed-effects as controls blocks this path and ensures we have conditional independence.
  2. **Additional controls increase precision:** Some variables (like a student's pre-test score) are strong predictors of the outcome but are uncorrelated with the random treatment. Including them in the regression explains a large part of the outcome's variance, which shrinks the standard errors on our treatment effect, giving us a more precise estimate.

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	Yes	Yes	Yes
R^2	.01	.25	.31	.31

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	—	—	—	−4.28 (1.96)
Male teacher	—	—	—	11.82 (3.33)
Teacher experience	—	—	—	.05 (0.06)
Master's degree	—	—	—	.48 (1.07)
School fixed effects	No	Yes	Yes	Yes
R^2	.02	.24	.30	.30

- 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

Emp. Challenges I: Threats to Internal Validity

- Even a perfectly designed RCT can face real-world challenges that threaten its internal validity. These issues can reintroduce the very biases that randomization was meant to solve.
- **Non-Random Attrition (Breaking the Randomization)**
  - **What it is:** Participants leave the experiment after being randomly assigned.
  - **The Threat:** If the reasons for leaving are related to the treatment or the outcome (e.g., the least motivated students in the large classes drop out), the remaining groups are no longer comparable. This breaks the independence assumption achieved by randomization and reintroduces selection bias.
  - **Example:** In the STAR experiment, if struggling students in large classes were more likely to move to private schools, a comparison of the remaining students would be biased and likely overstate the true effect of small classes.

Emp. Challenges I: Threats to Internal Validity

- **Imperfect Compliance (Breaking the Treatment Protocol)**
  - **What it is:** Participants do not adhere to their assigned group. Some in the treatment group may not take up the service ("no-shows"), while some in the control group may find a way to access it ("crossovers").
  - **The Threat:** A simple comparison of the original groups no longer measures the effect of receiving the treatment.
  - **The Solution:** This requires estimating two different effects:
    - Intention-to-Treat (ITT): Compare outcomes by the original assigned group, regardless of take-up. This estimates the causal effect of the policy offer.
    - Treatment-on-the-Treated (LATE): Use the initial random assignment as an instrumental variable (IV) for actual treatment received. This estimates the effect for the sub-group of "compliers."

Emp. Challenges I: Threats to Internal Validity

- **Example:** In Krueger (1999), students moved classes, see non-zero off-diagonals in transition matrix

B. First grade to second grade

	Second grade			
First 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

- **The Threat:** If students switched between TG and CG, then comparing them yields biased/inconsistent estimates
- **The Solution:** IV estimation, where the initial random assignment is the "instrument"

Emp. Challenges II: Threats to Interp. & Ext. Validity

Some challenges affect our interpretation the results and whether they apply to other contexts.

- **Behavioral Effects & Spillovers (Violating SUTVA)**
  - **What it is:** The experiment itself changes behavior in ways that contaminate the control group, violating the SUTVA "no interference" assumption.
  - **Examples:**
    - **Hawthorne Effect:** The treatment group changes behavior simply because they know they are being observed.
    - **John Henry Effect:** The control group works harder to compensate for not being in the treatment group.
    - **Substitution Bias:** Control group members seek out substitutes for the treatment they were denied. For example, parents of children in large classes might hire private tutors.
  - **The Threat:** The CG is no longer a clean counterfactual. The comparison is not between "treatment" and "no treatment", but between the TG and a CG whose behavior has also been altered. This will likely lead to an underestimation of the treatment effect.

Emp. Challenges II: Threats to Interp. & Ext. Validity

- **Heterogeneous Treatment Effects (Generalizability of the Effect)**
  - **What it is:** The impact of the treatment is not the same for everyone in the population.
  - **The Threat:** This is not a bias that invalidates the ATE for the experimental sample. Rather, it is a challenge for **external validity**.
    - If the people who volunteer for a trial are different from the general population (e.g., more motivated), the ATE found in the experiment may not reflect the ATE for the population as a whole.
    - Thus, the experimental result is a "local" effect valid for the participants

Emp. Challenges II: Threats to Interp. & Ext. Validity

- **Supply-Side & General Equilibrium Effects (Scaling Up)**
  - **What it is:** The effects measured in a small-scale pilot may not hold when the program is scaled up.
  - **Examples:**
    - **Supply-Side:** The teachers and staff in a pilot might be more motivated or better trained than those in a national roll-out.
    - **General Equilibrium:** A large-scale program can affect market prices. A massive job training program for welders could lower wages for all welders, affecting both treated and untreated individuals.
  - **The Threat:** These are fundamental threats to external validity. The results from the pilot may not accurately predict the impact of the scaled-up policy.

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 tesxt 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

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

Step 3: Substitute conditional means

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

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