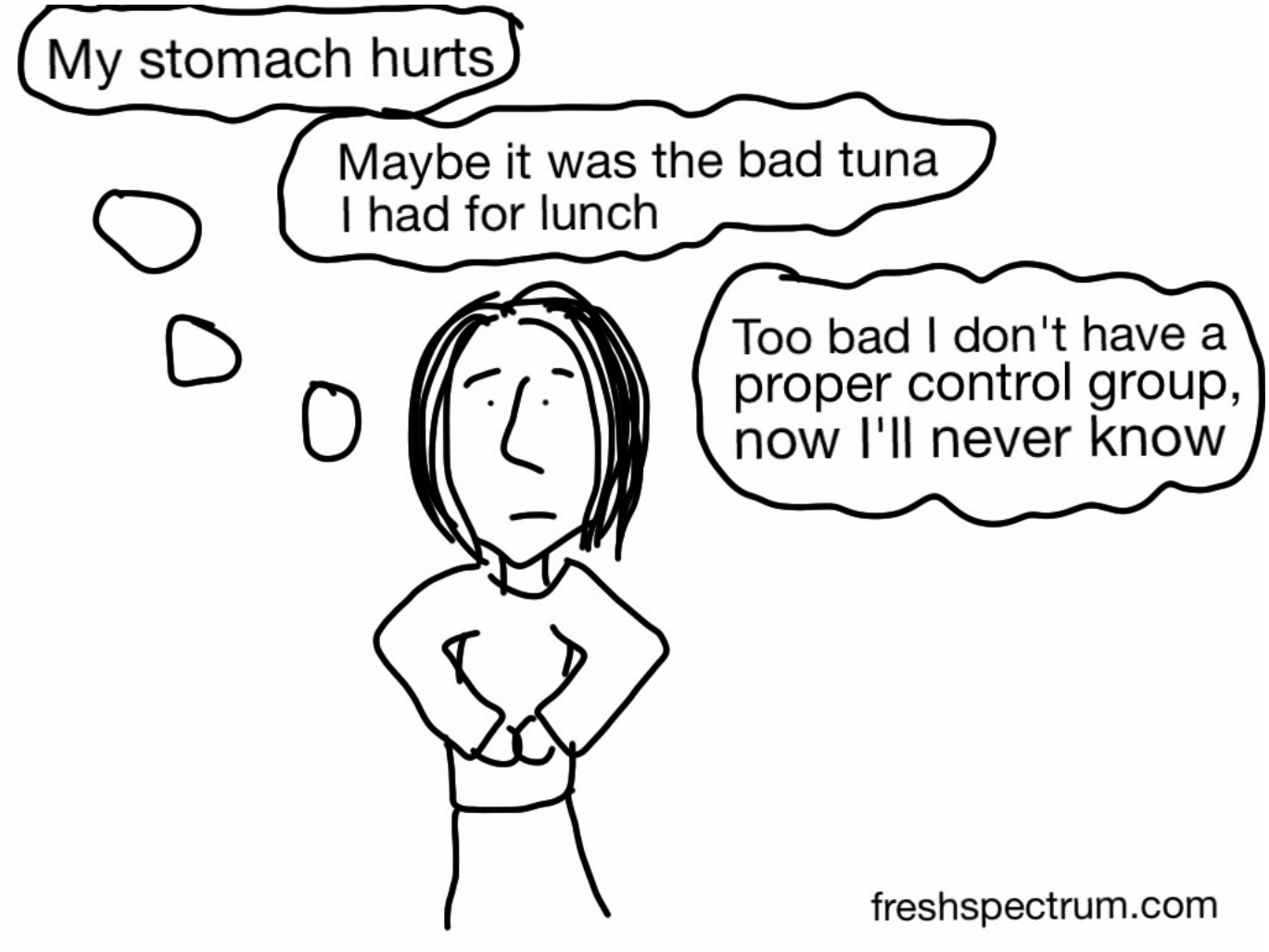
Lecture 5: Defining & Identifying Causal Relationships



Outline:

- Definition of potential outcomes
- Fundamental problem of causal inference
- Definition of causal parameters of interest: Average treatment effect (ATE) and average treatment effect on the treated (ATT)
- Moving from potential outcomes to regression
- Readings: A&P Chapters 1-2, I&W Lecture 1

Causal relationships:

- Basic concept of causality: an action (or 'treatment') T causes an outcome Y, if Y is the direct result of T
- 'Treatment effect' is the (average) causal effect of a binary (0–1) variable on an outcome variable of scientific or policy interest
  - The term 'treatment effect' comes from the medical literature concerned with the causal effects of binary, treatments, such as an experimental drug
- Examples of treatment: attending college, being exposed to air pollution, having an ITQ property rights regime, etc
  - Note: these are all binary treatments, i.e. T={0,1}. This is the case we consider here (most, but not all of what we will do translate to the case where 'treatment' is a dose, i.e. 0,1,2, …)

Notion of treatment:

- What is a treatment? Holland (JASA, 1986): a treatment is a potential manipulation that we can imagine. "No causation without manipulation"
  - Can I possibly expose each member of a population to T=0 or T=1?

- Q: Are gender, race, or adult height valid treatments? How about ambient pollution or smoking?

- Ex: You cannot manipulate race, but you can manipulate *perception* of race
  - Paper on this in next lecture

- Q: Any analogies to this point in natural sciences?
  - Can you manipulate climate? Topographical features?

Notion of treatment (ctd):

- The importance of the idea of “no causation without manipulation” is that it forces us think about what kind of questions we can possibly answer within the “causal inference” framework

- Specifically such questions are ones where we could conceptually imagine an ideal controlled experiment with random assignment of a ‘treatment’

- It should be noted that important questions are excluded from this approach. This does not mean they are not important (e.g., gender/racial discrimination)

Causality and potential outcomes

- We define a causal effect using a conceptual framework that is based on a set of *potential outcomes* that are observed in alternative states of the world

- "Rubin Causal Model", see Holland, JASA (1986)

- **Define:**
- T_i = **1**(unit i receives the treatment)

  $Y_i(0)$ = outcome of unit i realized if does not receive treatment

  $Y_i(1)$ = outcome of unit i realized if receives treatment

Potential outcomes: examples

- Importance: Potential outcomes are key building blocks of definitional and empirical models of causal effects

- Examples:

  $Y_i(0)$ = earnings of person i if she does not attend college

  $Y_i(1)$ = earnings of person i if she attends college

  $Y_i(0)$ = B_{MSY} of fishery i if ITQ system not implemented

  $Y_i(1)$ = B_{MSY} of fishery i if ITQ system is implemented

- *** Key challenge for any empirical investigation is that only one of ($Y_i(0)$, $Y_i(1)$) is observed for the same unit

Unit-level treatment effect:

- Denote $\beta_{1i} = Y_i(1) - Y_i(0)$ the causal effect of the treatment T_i for unit i (basic definition of causal effect)
  - Note: this is a theoretical construct since we do not observe $Y_i(1)$ and $Y_i(0)$ for the same i

- Important observations about β_{1i}:
  - Treatment effect heterogeneity is native to this model: the same treatment T_i may affect different units in a different way
  - Causal effects are defined independently of treatment assignment mechanism
  - Understanding how the treatment was assigned to subjects is key for causal inference

Unit-level treatment effect:

- Denote $\beta_{1i} = Y_i(1) - Y_i(0)$ the causal effect of the treatment T_i for unit i
  - Note: this is a theoretical construct since we do not observe $Y_i(1)$ and $Y_i(0)$ for the same i

- Important observations about β_{1i}:
  - No functional form assumption (i.e. not a linear regression or other type of model)
  - $\Rightarrow$ Definition of causality independent of estimation method
  - Goal is to estimate some feature of distribution of β_{1i} (typically a mean)

Fundamental problem of causal inference

- $Y_i(0)$, $Y_i(1)$ cannot be observed for the same unit during the same experiment
  - Because of this, we must rely on counterfactuals (assumed hypothetical unobserved value) to estimate causal effects
  - Need to estimate $Y_i(0)$ for treated, and $Y_i(1)$ for non-treated

- **Case 1**: Comparison of different units, exposed to different levels of the treatment

- **Case 2**: Comparison of the same units, at different points in time (for those exposed to different levels of a time-varying treatment). Clearly requires additional assumptions…
  - This is the basis for panel data analysis. “Not as general”

- $\Rightarrow$ The validity of the assumed counterfactual is the key to credible causal inference

SUTVA and observed outcomes

- SUTVA (Stable-Unit Treatment Value Assumption):
- Potential outcomes for unit i do not depend on the treatment assignments of other units (i.e. no interference across units)
- As a result, the observed outcome Y_i only depends on T_i:

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

- SUTVA is restrictive. It rules out: general equilibrium effects, peer effects, spillovers, interference, etc
- The alternative would be a model where the observed outcomes for unit i depend on the treatment received by the other units

Two Most Common Causal Parameters of Interest:

- **Average treatment effect (ATE):**

$$ATE = E[Y_i(1) - Y_i(0)]$$

  - ATE = average effect of the treatment for a randomly chosen unit

- **Average treatment effect on the treated (ATT):**

$$ATT = E[Y_i(1) - Y_i(0) \mid T_i = 1]$$

  - ATT= average treatment effect for units who received the treatment

- 1. In general ATE and ATT will differ. For example, if the units receiving the treatment are those benefiting from it the most on average, we will have ATT > ATE

- 2. ATE is an average on unconditional distributions of $Y_i(1)$ and $Y_i(0)$. When is that observed?

ATE and ATT in Context

- Example with fictional data set (red font unobserved)

i	$Y_i(1)$	$Y_i(0)$	β_{1i}	T_i	Y_i
1	5	2	3	1	5
2	1	1	0	1	1
3	1	0	1	0	0
4	1	1	0	0	1

ATE = $E[Y_i(1) - Y_i(0)] = 1$ and

ATT = $E[Y_i(1) - Y_i(0) \mid T_i = 1] = 1.5$ (notice that ATT > ATE)

Here $\overline{Y}_1 - \overline{Y}_0 = 2.5 \neq$ ATE $\neq$ ATT

- Key identification problem: in general you cannot identify ATE or ATT simply with data on Y_i and T_i.

Requirements for identification:

- Note that ATE can be written as:

$$\Pr(T_i=1)*E[Y_i(1) - Y_i(0) \mid T_i=1] + \{1-\Pr(T_i=1)\}*E[Y_i(1) - Y_i(0) \mid T_i=0]$$

- Thus for ATE, you need 2 counterfactuals, namely:

  $E[Y_i(0) \mid T_i=1]$ and $E[Y_i(1) \mid T_i=0]$

- For ATT, you only need 1 counterfactual: $E[Y_i(0) \mid T_i=1]$

- $\Rightarrow$ Stronger assumptions typically required to identify ATE as opposed to ATT

From potential outcomes to regression:

- Recall the Rubin Causal Model:

$Y_i(0)$ = potential outcome for unit i if untreated

$Y_i(1)$ = potential outcome for unit i if treated

T_i = 1 if unit i treated, 0 if not

- We observe (Y_i, T_i), where Y_i is the observed outcome:

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

$$= Y_i(0) + [Y_i(1) - Y_i(0)]T_i$$

$$= E[Y_i(0)] + [Y_i(1) - Y_i(0)]T_i + [Y_i(0) - E(Y_i(0))]$$

In words: we observe Y_i from the following conditional distributions: $Y_{1i}|T_i=1$ and $Y_{0i}|T_i=0$

From potential outcomes to regression (ctd):

- Thus:

$$Y_i = E[Y_i(0)] + [Y_i(1) - Y_i(0)]T_i + [Y_i(0) - E(Y_i(0))]$$
$$= \beta_0 + \beta_{1i}T_i + u_i$$

- Where:

$\beta_0 = E[Y_i(0)]$

$\beta_{1i} = Y_i(1) - Y_i(0)$ = individual treatment effect

$u_i = Y_i(0) - E(Y_i(0))$, so $E(u_i) = 0$

Causal parameters in this regression framework:

- Causal parameters in this framework:

$\beta_{1i} = Y_i(1) - Y_i(0)$ = individual treatment effect

$E[\beta_{1i}]$ = average treatment effect (ATE)

$E[\beta_{1i} | T_i = 1]$ = average treatment effect on the treated (ATT)

- When there is a single treatment effect (i.e. constant treatment effect) then $\beta_{1i} = \beta_1$, we obtain the usual regression model:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

- where β_1 has causal interpretation = average treatment effect parameter (which we assume here to be constant)

Identification of ATE in this framework:

- The model of potential outcomes, with constant treatment effects implies the standard regression model:

$$Y_i = \beta_0 + \beta_1 T_i + u_i$$

- Can we correctly estimate the causal effect of T_i on Y_i with data on T_i and Y_i?
  - Can we estimate regression coefficient β_1 without bias?
- The key is whether the assumption $Cov(T_i, u_i) = 0$ is satisfied [u and T uncorrelated]
  - In other words, is LSA#1 satisfied
  - When is this a reasonable assumption?
- Later: Address treatment effect heterogeneity ($\beta_{1i} \neq \beta_1$)

Omitted variables bias in this simple model:

- If T_i and u_i are uncorrelated (maybe conditional on other characteristics X_i), then we proceed with OLS to estimate average causal effect of T_i on Y_i:
  - The key to support this approach will be to identify research strategy (treatment assignment mechanism) where the LSA1 assumption above (or a variant) is credibly satisfied

- However, if T_i and u_i are correlated, the OLS is biased due to omitted variables bias:

$$\text{plim}\,\hat{\beta}_1^{OLS} = \frac{Cov(Y_i, T_i)}{Var(T_i)}$$

$$= \frac{Cov(\beta_0 + \beta_1 T_i + u_i, T_i)}{Var(T_i)} = \beta_1 + \frac{\text{Cov}(u_i, T_i)}{\text{Var}(T_i)}$$

Three Treatment Assignment Mechanisms:

- 1. Random assignment of treatment (ensures that treatment is independent of potential outcomes)
  - Often the “ideal” benchmark to contrast with other studies

- 2. Treatment not independent of potential outcomes, selection into T={0,1} based only on observable variables:
  - "Selection on observables” ⇒ can "control" for selection if you know the selection rule and have the appropriate data

- 3. Treatment not independent of potential outcomes, selection into T={0,1} based on unobservable variables:
  - "Selection on unobservables" ⇒ Need valid instruments, regression discontinuity, panel data, or need to know the statistical distribution of the unobservables