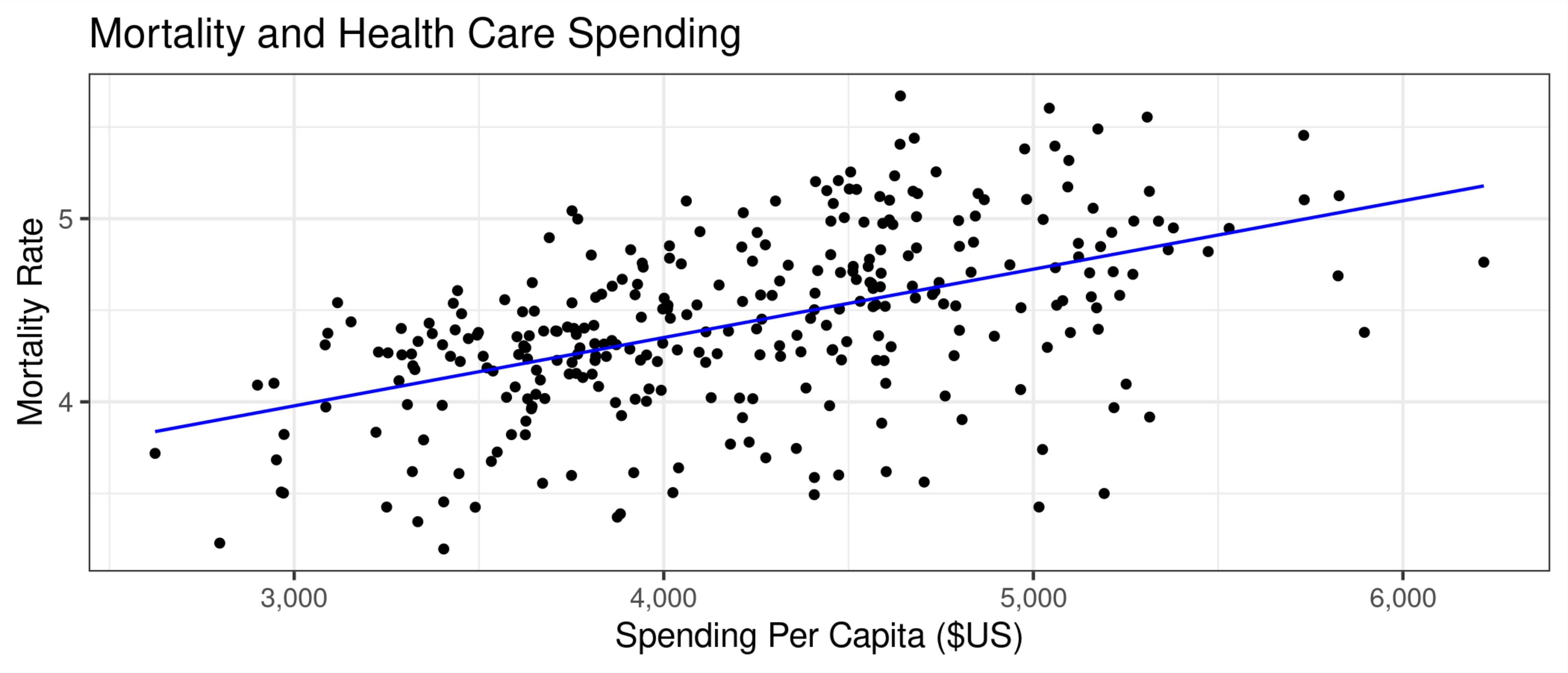
Module 1: Hospital Pricing and Selection on Observables

Part 2: Introduction to Causal Inference

Ian McCarthy | Emory University

Econ 470 & HLTH 470

Why causal inference?



Why causal inference?

Another example: **What price should we charge for a night in a hotel?**

Machine Learning

- Focuses on prediction
- High prices are strongly correlated with higher sales
- Increase prices to attract more people?

Causal Inference

- Focuses on **counterfactuals**
- What would sales look like if prices were higher?

Goal of Causal Inference

- **Goal:** Estimate effect of some policy or program
- Key building block for causal inference is the idea of **potential outcomes**

Some notation

Treatment D_i

$$D_i = \begin{cases} 1 \text{ with treatment} \\ 0 \text{ without treatment} \end{cases}$$

Some notation

Potential outcomes

- Y_{1i} is the potential outcome for unit i with treatment
- Y_{0i} is the potential outcome for unit i without treatment

Some notation

Observed outcome

$$Y_i = Y_{1i} \times D_i + Y_{0i} \times (1 - D_i)$$

or

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases}$$

Assumes **SUTVA** (stable unit treatment value assumption)...no interference across units

Example of "Potential Outcomes"

Y_1 = \$75,000

Y_0 = \$60,000

Example of "Potential Outcomes"

Y_1 = \$75,000

Y_0 = \$60,000

Earnings due to Emory = $Y_1 - Y_0$ = \$15,000

Example of "Potential Outcomes"

Y_1 = $75,000

Y_0 = ?

Example of "Potential Outcomes"

Y_1 = \$75,000

Y_0 = ?

Earnings due to Emory = $Y_1 - Y_0$ = ?

Do we ever observe the potential outcomes?

Without a time machine…not possible to get *individual* effects.

Fundamental Problem of Causal Inference

- We don't observe the counterfactual outcome...what would have happened if a treated unit was actually untreated.

- *ALL* attempts at causal inference represent some attempt at estimating the counterfactual outcome. We need an estimate for Y_0 among those that were treated, and vice versa for Y_1.

Average Treatment Effects

Different treatment effects

Tend to focus on **averages**[1]:

- **ATE**: $\delta_{ATE} = E[Y_1 - Y_0]$
- **ATT**: $\delta_{ATT} = E[Y_1 - Y_0|D = 1]$
- **ATU**: $\delta_{ATU} = E[Y_1 - Y_0|D = 0]$

[1] or similar measures such as medians or quantiles

Average Treatment Effects

- **Estimand**:

$$\delta_{ATE} = E[Y_1 - Y_0] = E[Y|D=1] - E[Y|D=0]$$

- **Estimate**:

$$\hat{\delta}_{ATE} = \frac{1}{N_1} \sum_{D_i=1} Y_i - \frac{1}{N_0} \sum_{D_i=0} Y_i,$$

  where N_1 is number of treated and N_0 is number untreated (control)

- With random assignment and equal groups, inference/hypothesis testing with standard two-sample t-test

Selection Bias

Selection bias

- Assume (for simplicity) constant effects, $Y_{1i} = Y_{0i} + \delta$
- Since we don't observe Y_0 and Y_1, we have to use the observed outcomes, Y_i

$$
\begin{aligned}
E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\
&= E[Y_{1i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&= \delta + E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \\
&= \text{ATE} + \text{ Selection Bias}
\end{aligned}
$$

Selection bias

- Selection bias means $E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] \neq 0$
- In words, the potential outcome without treatment, Y_{0i}, is different between those that ultimately did and did not receive treatment.
- e.g., treated group was going to be better on average even without treatment (higher wages, healthier, etc.)

Selection bias

- How to "remove" selection bias?

- How about random assignment?

- In this case, treatment assignment doesn't tell us anything about Y_{0i}

$$E[Y_{0i}|D_i = 1] = E[Y_{0i}|D_i = 0],$$

such that

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = \delta_{ATE} = \delta_{ATT} = \delta_{ATU}$$

Selection bias

- Without random assignment, there's a high probability that

$$E[Y_{0i}|D_i = 1] \neq E[Y_{0i}|D_i = 0]$$

- i.e., outcomes without treatment are different for the treated group

Omitted variables bias

- In a regression setting, selection bias is the same problem as omitted variables bias (OVB)
- Quick review: Goal of OLS is to find $\hat{\beta}$ to "best fit" the linear equation $\boldsymbol{y_i = \alpha + x_i \beta + \epsilon_i}$

Regression review

$$\min_{\beta} \sum_{i=1}^{N} (y_i - \alpha - x_i\beta)^2 = \min_{\beta} \sum_{i=1}^{N} (y_i - (\bar{y} - \bar{x}\beta) - x_i\beta)^2$$

$$0 = \sum_{i=1}^{N} \left(y_i - \bar{y} - (x_i - \bar{x})\hat{\beta}\right)(x_i - \bar{x})$$

$$0 = \sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x}) - \hat{\beta} \sum_{i=1}^{N}(x_i - \bar{x})^2$$

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = \frac{Cov(y, x)}{Var(x)}$$

Omitted variables bias

- Interested in estimate of the effect of schooling on wages

$$Y_i = \alpha + \beta s_i + \gamma A_i + \epsilon_i$$

- But we don't observe ability, A_i, so we estimate

$$Y_i = \alpha + \beta s_i + u_i$$

- What is our estimate of β from this regression?

Omitted variables bias

$$
\begin{aligned}
\hat{\beta} &= \frac{Cov(Y_i, s_i)}{Var(s_i)} \\
&= \frac{Cov(\alpha + \beta s_i + \gamma A_i + \epsilon_i, s_i)}{Var(s_i)} \\
&= \frac{\beta Cov(s_i, s_i) + \gamma Cov(A_i, s_i) + Cov(\epsilon_i, s_i)}{Var(s_i)} \\
&= \beta \frac{Var(s_i)}{Var(s_i)} + \gamma \frac{Cov(A_i, s_i)}{Var(s_i)} + 0 \\
&= \beta + \gamma \times \theta_{as}
\end{aligned}
$$

Removing selection bias without RCT

- The field of causal inference is all about different strategies to remove selection bias
- The first strategy (really, assumption) in this class: **selection on observables** or **conditional indpendence**

Intuition

- Example: Does having health insurance, $D_i = 1$, improve your health relative to someone without health insurance, $D_i = 0$?
- Y_{1i} denotes health with insurance, and Y_{0i} health without insurance (these are **potential** outcomes)
- In raw data, $[Y_i|D_i = 1] > E[Y_i|D_i = 0]$, but is that causal?

Intuition

Some assumptions:

- $Y_{0i} = \alpha + \eta_i$
- $Y_{1i} - Y_{0i} = \delta$
- There is some set of "controls", x_i, such that $\eta_i = \beta x_i + u_i$ and $E[u_i|x_i] = 0$ (conditional independence assumption, or CIA)

$$\begin{aligned}
Y_i &= Y_{1i} \times D_i + Y_{0i} \times (1 - D_i) \\
&= \delta D_i + Y_{0i} D_i + Y_{0i} - Y_{0i} D_i \\
&= \delta D_i + \alpha + \eta_i \\
&= \delta D_i + \alpha + \beta x_i + u_i
\end{aligned}$$

ATEs versus regression coefficients

- Estimating the regression equation,

$$Y_i = \alpha + \delta D_i + \beta x_i + u_i$$

  provides a causal estimate of the effect of D_i on Y_i

- But what does that really mean?

ATEs vs regression coefficients

- *Ceteris paribus* ("with other conditions remaining the same"), a change in D_i will lead to a change in Y_i in the amount of $\hat{\delta}$
- But is *ceteris paribus* informative about policy?

ATEs vs regression coefficients

- $Y_{1i} = Y_{0i} + \delta_i D_i$ (allows for heterogeneous effects)
- $Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$, with $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$
- Aronow and Samii, 2016, show that:

$$\hat{\beta} \rightarrow_p \frac{E[w_i \delta_i]}{E[w_i]},$$

where $w_i = (D_i - E[D_i|X_i])^2$

ATEs vs regression coefficients

- Simplify to ATT and ATU
- $Y_{1i} = Y_{0i} + \delta_{ATT} D_i + \delta_{ATU}(1 - D_i)$
- $Y_i = \alpha + \beta D_i + \gamma X_i + \epsilon_i$, with $Y_{0i}, Y_{1i} \perp\!\!\!\perp D_i | X_i$

$$\beta = \frac{P(D_i = 1) \times \pi(X_i|D_i = 1) \times (1 - \pi(X_i|D_i = 1))}{\sum_{j=0,1} P(D_i = j) \times \pi(X_i|D_i = j) \times (1 - \pi(X_i|D_i = j))} \delta_{ATU}$$
$$+ \frac{P(D_i = 0) \times \pi(X_i|D_i = 0) \times (1 - \pi(X_i|D_i = 0))}{\sum_{j=0,1} P(D_i = j) \times \pi(X_i|D_i = j) \times (1 - \pi(X_i|D_i = j))} \delta_{ATT}$$

ATEs vs regression coefficients

What does this mean?

- OLS puts more weight on observations with treatment D_i "unexplained" by X_i
- "Reverse" weighting such that the proportion of treated units are used to weight the ATU while the proportion of untreated units enter the weights of the ATT
- This is *an* average effect, but probably not the average we want