Outline

1) Causality

2) Potential Outcomes

3) Randomized Experiments

Causality

Correlation $\neq$ Causation

So far, course has focused on drawing
inferences about probability dist.s

<u>Ex</u> $D_i = 1\{\text{Student } i \text{ gets scholarship}\}$
Y_i = College GPA of student i

Given iid sample of (D_i, Y_i) pairs, could
- estimate $\mathbb{E}[Y_i \mid D_i]$
- get interval for $\theta = \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$
- test $H_0 : \text{dist}(Y_i \mid D_i = 1) = \text{dist}(Y_i \mid D_i = 0)$

<u>But</u> Suppose CI for θ is $[0.7, 0.8]$
Still can't conclude scholarship <u>caused</u> GPA boost

(Why not?)

People often interpret regression coeff.s causally,
"all other things equal" – usually w/o justification

We can never draw causal conclusions from joint dist. <u>alone</u>

<u>But</u> causal questions are often very interesting, drive policy, etc.

When can we draw causal conclusions?

Two questions about causality:

<u>Hard question</u>: what caused an outcome?

"Why couldn't he find a job?"

- very few things are monocausal
- causes may interact w/ each other

<u>Easier</u>: what effect does an intervention have?

"Does a specific job training program help people find jobs?"

- more tractable — could imagine experiment
- only meaningful relative to specific (control) counterfactual (same person, no training program)

"Effect of a cause" not "Cause of an effect"

Potential Outcomes

Idea: Define causal effect in terms of outcomes that *would* happen in alt. universe

Ex (Cont'd):

$Y_i(1)$ = GPA of student i with schol.

$Y_i(0)$ = GPA " w/o schol.

Observed $Y_i = Y_i(D_i)$ (SUTVA)

(Implicit: Y_i only depends on D_i (not D_j for $j \neq i$))

Problem: Only one potential outcome observed

Define average treatment effect (ATE)

as $\delta = \mathbb{E}[Y_i(1) - Y_i(0)]$

Different from θ:

$$\theta = \mathbb{E}[Y_i(1) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]$$

$$= \mathbb{E}[Y_i(1) - Y_i(0) \mid D_i = 1] + \underbrace{\mathbb{E}[Y_i(0) \mid D_i = 1] - \mathbb{E}[Y_i(0) \mid D_i = 0]}$$

ATT

selection bias

Randomized Experiments

Suppose D_i assigned at random,
w/o regard to any attributes

Then, $\mathbb{E}[Y_i(1) \mid D_i = 1] = \mathbb{E}\, Y_i(1)$

$\Rightarrow \theta = \delta$

Can use usual two-sample methods,
interpret causally.

For example, can get unbiased
estimate of $\delta = \theta$:

$$\hat{\delta} = \frac{1}{n_1} \sum_{i: D_i = 1} Y_i - \frac{1}{n_0} \sum_{i: D_i = 0} Y_i$$

Usually can't observe both potential outcomes
in social science applications

Randomization Tests

What if we want to test whether treatment has <u>any</u> effect (on any unit)?

<u>Fisher's sharp null:</u>

$$H_0: \quad Y_i(1) = Y_i(0) \qquad \forall i = 1, \dots, n$$

<u>Note:</u> statement about (unobserved) aspect of these n units (no sampling model)

<u>Idea:</u> Use randomness of treatment assignment to our benefit

Under H_0, $Y_i(D_i^*) = Y_i(D_i)$ under <u>any</u> treatment assignment

We know dist. of $D \Rightarrow$ know (null) dist of any test stat $T(D, Y(D))$

For $b = 1, \ldots, B$:

$$D^{*b} = \text{permute}(D)$$

$$\hat{\delta}^{*b} = \frac{1}{n_1} \sum_{i: D_i^{*b} = 1} Y_i - \frac{1}{n_0} \sum_{i: D_i^{*b} = 0} Y_i$$

Under H_0, $\hat{\delta}, \hat{\delta}^{*1}, \ldots, \hat{\delta}^{*B}$ are exchangeable $\Rightarrow$ use $p = \frac{1}{B+1}\left(1 + \#\{\hat{\delta}^{*b} \geq \hat{\delta}\}\right)$

Can extend to CI for constant treatment effect:

$$H_0^{\delta}: \; Y_i(1) - Y_i(0) = \delta \quad \forall i$$

Under H_0, $Y_i(D_i^{*b}) - Y_i(D_i) = \delta(D_i^{*b} - D_i)$

$$T^{*b} = \frac{1}{n_1} \sum_{D_i^{*b} = 1} \left(Y_i + \delta(1 - D_i)\right) - \frac{1}{n_0} \sum_{D_i^{*b} = 0} \left(Y_i - \delta D_i\right)$$

Experimental Design

Often can improve precision of estimator if we randomize more carefully

e.g. Suppose pre-treatment covariate X_i is highly predictive of outcome

say, X_i = age

$Y_i = 1\{\text{recovers from illness}\}$

If treatment assigned unif. randomly, treatment group will get more or less older people by chance

Could be main driver of variance!

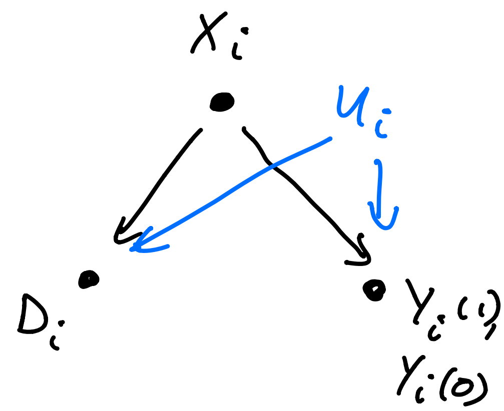
Better: Match 2 oldest, next 2, ..., 2 youngest

Randomize w/in pairs

$\rightsquigarrow$ Different randomization test.

Unconfoundedness

In observational studies, we don't randomize, just observe world as it is.



Assume $(X_i, D_i, Y_i(0), Y_i(1)) \overset{iid}{\sim} P$

pre-treatment covariate ↗ treatment ↑ P.O.s ↖

we want $\delta = \mathbb{E}_P[Y_i(1) - Y_i(0)]$

Suppose $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i \mid X_i$

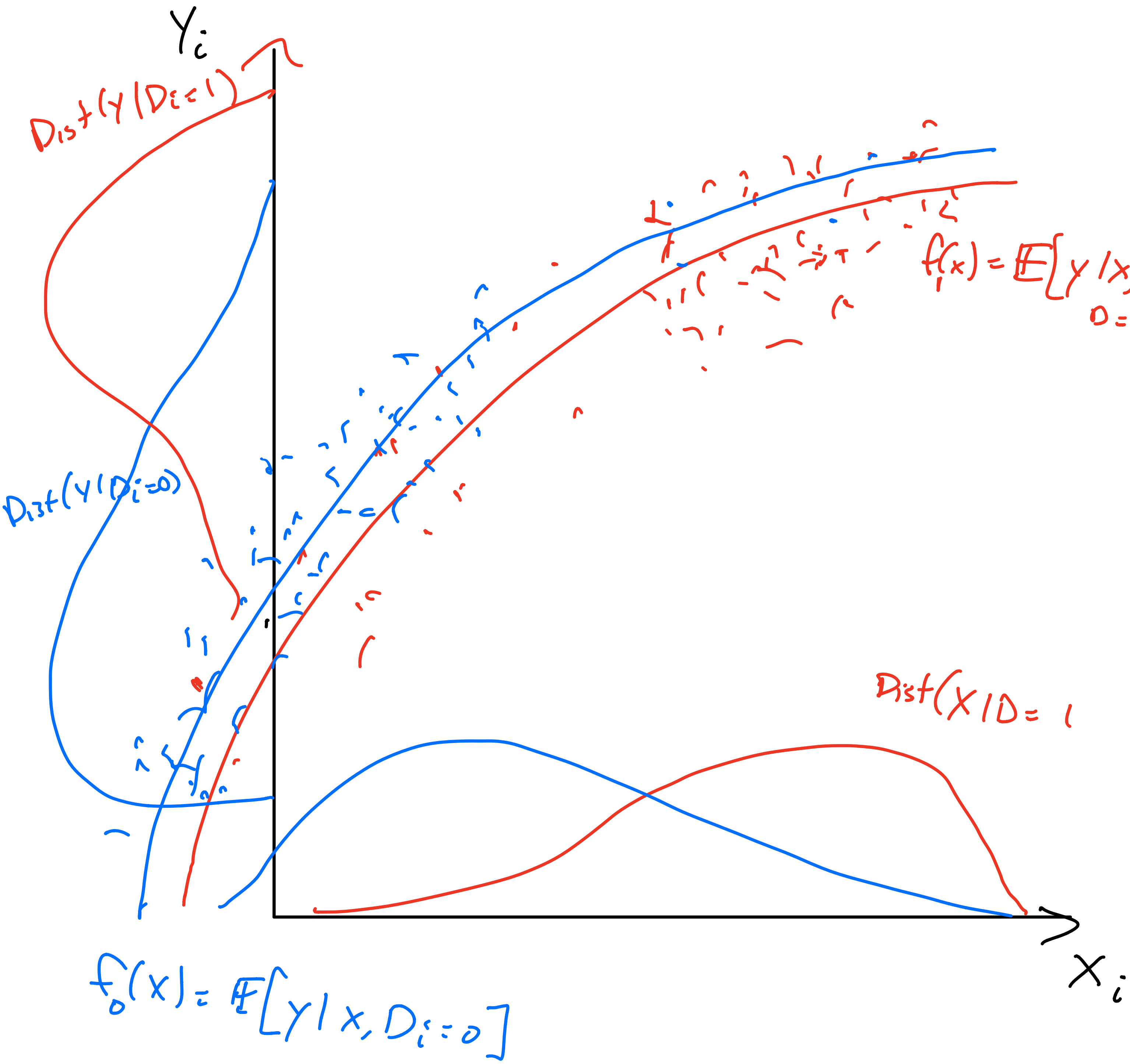
then δ becomes identifiable from obs. dist.

$$\mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\mathbb{E}[Y_i(1) \mid X_i] - \mathbb{E}[Y_i(0) \mid X_i]]$$

$$= \mathbb{E}\{\mathbb{E}\{\underbrace{Y_i}_{\text{observed } Y_i} \mid X_i, D_i = 1\} - \mathbb{E}\{Y_i \mid X_i, D_i = 0\}\}$$

$$= \mathbb{E}[f_1(X_i) - f_0(X_i)]$$

$f_d(X_i) = \mathbb{E}[Y_i \mid X_i, D_i = d]$ identifiable.



Example of Simpson's Paradox: apparent "effect" is reversed after we condition on something.

Propensity Scores

Sometimes can't randomize ourselves, but <u>can</u> estimate how "nature" randomized treatment assignment.

Suppose we observe covariates X_i, know

$$e(x) = \mathbb{P}(D_i = 1 \mid X_i = x)$$

Propensity score function

$(e(x) \in (0,1))$

And assume <u>unconfoundedness</u>:

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid X_i$$

Then also true that $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i \mid e(X_i)$

Why?

$$\mathbb{P}(Y_i \in A, D_i = 1 \mid e(X_i)) = \mathbb{E}\{ \mathbb{P}(Y_i \in A, D_i = 1 \mid X_i) \mid e(X_i) \}$$

$$= \mathbb{E}\{ \mathbb{P}(Y_i \in A \mid X_i) \overbrace{\mathbb{P}(D_i = 1 \mid X_i)}^{= e(X_i)} \mid e(X_i) \}$$

$$= e(X_i) \, \mathbb{E}\{ \mathbb{P}(Y_i \in A \mid X_i) \mid e(X_i) \}$$

$$= e(X_i) \, \mathbb{P}(Y_i \in A \mid e(X_i))$$

We could use regression adjustment

$$\mu_d = \mathbb{E}[\mathbb{E}\{ Y_i \mid e(X_i), D_i = d \}]$$

But can actually make exact adj. instead.

Inverse Propensity Weighting

Suppose we actually know

$$e(x) = \mathbb{P}(D_i = 1 \mid X_i = x)$$

Could use it to consistently est. τ:

$$\hat{\tau} = \frac{1}{n} \sum_{D_i=1} \frac{Y_i}{e(X_i)} - \frac{1}{n} \sum_{D_i=0} \frac{Y_i}{1-e(X_i)}$$

not n_1

Informally: for each $(X_i, Y_i(0), Y_i(1))$ we have

an $e(X_i)$ chance of observing $Y_i(1)$,

(otherwise $D_i = 0$ and $Y_i(1)$ missing).

$\Longrightarrow$ upweight observed ones accordingly

(works as long as $e(x) \in (0,1)$ $\forall x$)

More formal:

$$\frac{1}{n}\sum_{D_i=1}\frac{Y_i}{e(X_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(1)D_i}{e(X_i)}$$

$$\mathbb{E}\left[\frac{Y_i(1)D_i}{e(X_i)} \,\Big|\, X_i\right] = \mathbb{E}\left[Y_i(1) \mid X_i\right]\cdot \underbrace{\mathbb{E}\left[\frac{D_i}{e(X_i)} \,\Big|\, X_i\right]}_{=1}$$

$$\Rightarrow \mathbb{E}\frac{Y_i D_i}{e(X_i)} = \mathbb{E}\, Y_i(1)$$

Similarly, $\displaystyle \frac{1}{n}\sum_{D_i=0}\frac{Y_i}{1-e(X_i)} = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i(0)(1-D_i)}{1-e(X_i)}$

$$\Rightarrow \mathbb{E}\frac{Y_i(1-D_i)}{1-e(X_i)} = \mathbb{E}\, Y_i(0)$$

So, $\mathbb{E}\,\hat{\delta} = \mathbb{E}\left(Y_i(1) - Y_i(0)\right) = \delta$

<u>Practical issues:</u> if $e(X_i) \approx 0$ or ≈ 1 for some X_i, IPW estimator can have extremely high variance. Called "poor overlap"

Estimated Prop. Scores

Problem: we don't know $e(X_i)$

Could estimate, e.g. using logistic regression

Different e ☹

$$e(x) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

Use $$\hat{e}(x) = \frac{e^{\hat{\beta}' x}}{1 + e^{\hat{\beta}' x}}$$

Bias $\to 0$, $\hat{\delta}$ still consistent

If propensity score model is misspecified,

Bias $\not\to 0$, $\hat{\delta}$ not consistent

Unsurprising since our model was wrong

<u>More surprising</u>: we can (sometimes) correct it!

What is the bias? Assume f_0

Write $\hat{\mu}_1 = \frac{1}{n} \sum \frac{Y_i D_i}{e(X_i)}$, $\hat{\mu}_0 = \frac{1}{n} \sum \frac{Y_i (1-D_i)}{1-e(X_i)}$

then $\hat{\delta} = \hat{\mu}_1 - \hat{\mu}_0$

$$\text{Bias}(\hat{\mu}_1) = \mathbb{E}\left[\frac{Y_i(1) D_i}{\hat{e}(X_i)} - Y_i(1)\right]$$

Assume for simplicity $\hat{e}(\cdot) \perp\!\!\!\perp (X_i, D_i, Y_i)_{i=1}^n$

(e.g. data splitting)

$$\mathbb{E}\left[\frac{Y_i(1) D_i}{\hat{e}(X_i)} - Y_i(1) \mid X_i\right] = \mathbb{E}\left[Y_i(1) \cdot \frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)} \mid X_i\right]$$

$$= \overbrace{\mathbb{E}\left[Y_i(1) \mid X_i\right]}^{f_1(X_i)} \mathbb{E}\left[\frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)} \mid X_i\right]$$

$$\Longrightarrow \text{Bias}(\hat{\mu}_1) = \mathbb{E}\left[f_1(X_i) \cdot \frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)}\right]$$

Similarly, $\text{Bias}(\hat{\mu}_0) = \mathbb{E}\left[f_0(X_i) \cdot \frac{\hat{e}(X_i) - D_i}{1 - e(X_i)}\right]$

We can estimate/correct Bias using regression!

Double Robustness

(Assume for simplicity $\hat{f}_0, \hat{f}_1, \hat{e} \perp\!\!\!\perp$ data)

Let $\hat{\delta}^{DR} = \hat{\mu}_1^{DR} - \hat{\mu}_0^{DR}$, where

$$\hat{\mu}_1^{DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i D_i}{\hat{e}(X_i)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{D_i - \hat{e}(X_i)}{\hat{e}(X_i)} \hat{f}_1(X_i)}_{\widehat{\text{Bias}}(\hat{\mu}_1)}$$

$$\hat{\mu}_0^{DR} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i(1-D_i)}{1-\hat{e}(X_i)} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \frac{\hat{e}(X_i) - D_i}{1-\hat{e}(X_i)} \hat{f}_0(X_i)}_{\widehat{\text{Bias}}(\hat{\mu}_0)}$$

<u>Claim</u> If <u>either</u>

- $\hat{e}(x) = e(x)$ $(+\ o_p(1))$ <u>or</u>
- $\hat{f}_d(x) = f_d(x)$ $(+\ o_p(1))$, for $d=0,1$

Then $\mathbb{E}\,\hat{\delta}^{DR} = \delta$ $(+\ o_p(1))$

Proof

1. Suppose $\hat{f}_d(x) = f_d(x)$ for $d = 0, 1$

By previous derivation,

$$\mathbb{E}\, \hat{\mu}_d^{DR} = \mathbb{E}\, \hat{\mu}_d - \text{Bias}(\hat{\mu}_d) = \mu_d$$

2. Suppose $\hat{e}(x) = e(x)$, $\hat{f}_0, \hat{f}_1$ possibly wrong

$$\mathbb{E}\left[\hat{f}_1(x_i) \frac{D_i - \hat{e}(x_i)}{\hat{e}(x_i)} \mid X_i \right]$$

$$= \frac{\hat{f}_1(x_i)}{\hat{e}(X_i)} \cdot \xcancel{\mathbb{E}\left[D_i - \hat{e}(x_i) \mid X_i \right]} \rightarrow 0$$

So $\mathbb{E}\, \hat{\mu}_1^{DR} = \mathbb{E}\, \hat{\mu}_1 = \mu_1$

Similar for $\hat{\mu}_0^{DR}$