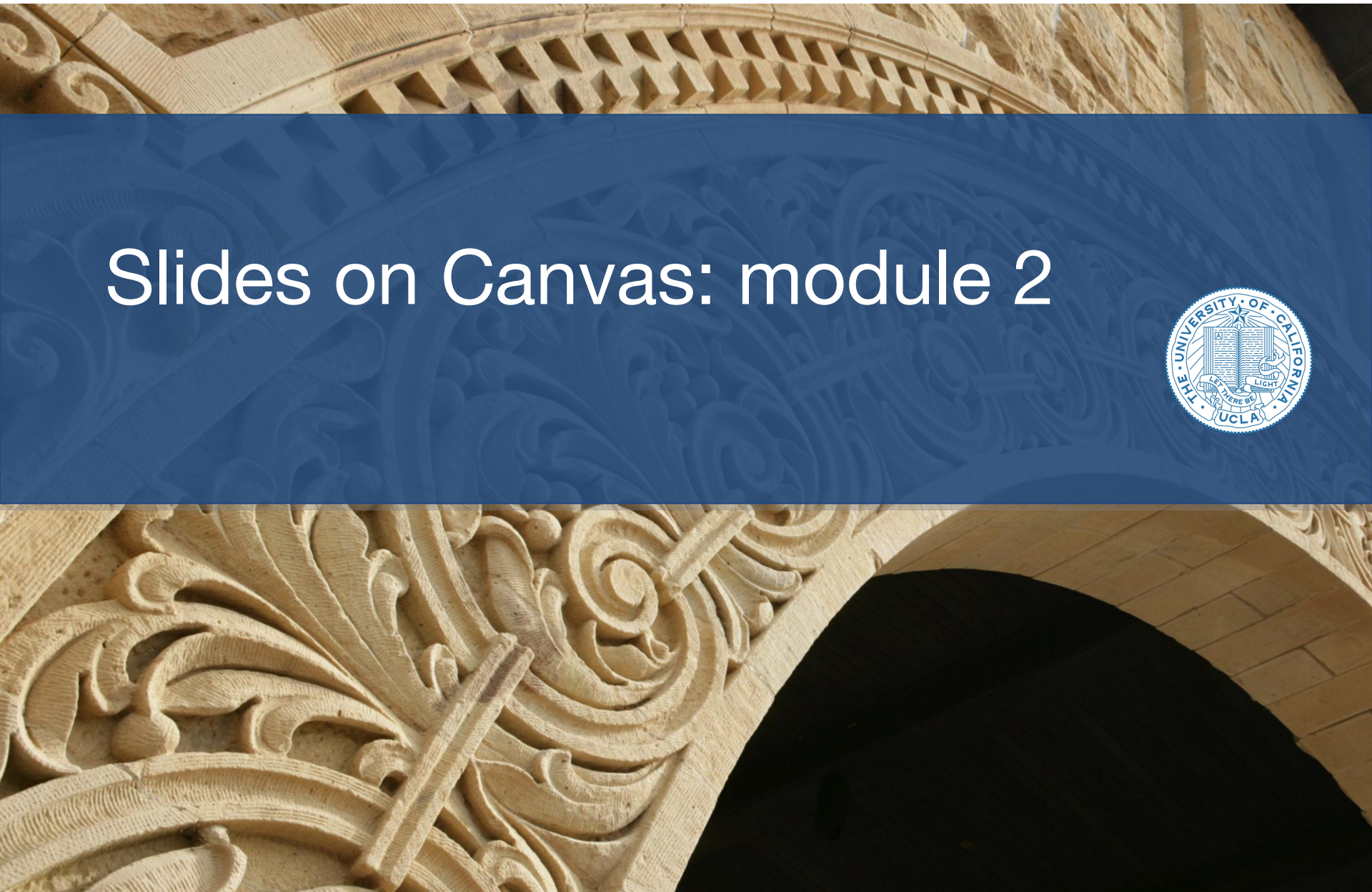
UCLAAnderson
SCHOOL of MANAGEMENT

PMDA

Lecture 2: Causality & A/B testing

UCLAAnderson
SCHOOL of MANAGEMENT

Slides on Canvas: module 2



Housekeeping

- TA office hours:
  - Section 1: Thursdays, 3 PM with Jian
    - Zoom Meeting: https://ucla.zoom.us/j/91810277540
  - Section 2: Tuesdays, 5:00 PM with Martin
    - Zoom Meeting: https://ucla.zoom.us/j/96914285879
- Problem Set # 1: Jan 27, 11.59 pm

Recap from last class

- Interpreting univariate regression output:
  - X continuous: expected change in Y for a one unit change in X
  - X discrete: difference in means between group with X=1 and group with X=0 (regression is logically equivalent to 2-sample mean test)

- Precision and hypothesis testing:
  - Standard error = uncertainty of coefficient estimate
  - T-stat: Under hypothesis of no effect, how far away from zero is the estimated coefficient
  - P-value: Under null of no effect, probability of observing an effect as far away from zero as the one estimated (reject if p-value<0.05)

Road map

Agenda

1. Causality & correlation
2. Omitted variable bias
3. A/B testing & causality

Goals

- Understand difference between correlation & causation
- Understand omitted variable bias
  - When it arises
  - What determines the direction of the bias
  - How to fix it
- Interpret coefficients in multivariate regression
- Understand why A/B tests give causal estimates

Observational vs. Experimental Studies

- To understand causation vs. correlation in empirical studies it is important to distinguish between two types of studies:
    - Observational studies/data = where we simply measure what people do without trying to affect their behavior
    - Experimental studies/data = data from an experiment that affects behavior (also called A/B testing or Randomized Controlled Trial)
- Hard to distinguish causation from correlation in observational studies because of the problem of endogeneity
- Experimental studies do not suffer from endogeneity and thus give you causation
- Let's start with typical examples of observational studies

Example (1): Red Meat Consumption

- **Q:** Recall from last lecture that in survey data (= observational study) red meat is correlated with cardiac risk. Is this causation?

- **A:** We saw that in the data people who eat red meat tend to also smoke, so you may be picking up the effect of smoking too

→Here you worry that correlation is not causation (endogeneity problem) because of omitted variables (smoking)

Example (2): Value of your UCLA Degree

- What is the causal effect of studying at UCLA (rather than USC) on your future earnings?

- Say you have observational data on graduate earnings and compare the earnings of graduates from both schools

- **Q:** What is the problem with this?

- **A:** Schools might admit different types of students, i.e. if UCLA is more selective or if smarter students choose UCLA

→Here you worry that correlation is not causation (endogeneity problem) because of omitted variables (students' quality/ ability)

Example (3): Policing & Crime

- How effective is policing in deterring crime?
- Say we have neighborhood-level data on the number of police officers (X) deployed and the level of crime (Y)
- **Q:** Would you expect a positive or negative correlation between X and Y?
- **A:** Often positive because more police officers get deployed to areas with more crime. This is a classic case of Y causing X more than X causing Y

→Here you worry that correlation is not causation (endogeneity problem) because of reverse causality (a.k.a. simultaneity bias)

Causality & why we care about it

Causality:

- The impact of “X and X alone” on Y
- Usually because we are interested in the question: if we could manipulate only X, how would it affect Y?
    - Effect on cardiac risk of becoming vegetarian
    - Effect on earnings of moving a USC student to UCLA
    - Effect on crime of increasing policing

Fundamental problem: in observational data X typically does not vary independently of other factors

- Makes it difficult to learn about the impact of X alone
- Instead we only learn about correlation

Note that correlations can be useful for prediction…

- Example: *In data from fire department deployments, we find that when there are more fire trucks, the fire damage tends to be larger*

- **Q:** Do fire trucks cause larger fires?
- **A:** Causality clearly runs the other way here. Larger fires lead to more trucks

- **Q:** Can it be useful to know the number of trucks?
- **A:** If you don't know anything else about the fire, then knowing the number of trucks can help you assess the magnitude of the fire

→Correlations are useful for prediction, but not for understanding what happens when we manipulate X (in this case: number of trucks)

Causality: A simple framework

- Let's assume we observe correlation in observational data. We need logic to establish a causal link

- 3 possible reasons why X and Y are correlated:

1. (Effect of interest) X causes Y
2. (Reverse causality) Y causes X
3. (Omitted variables) There is a variable Z that causes X and Y (Z causes Y and is correlated with X or Z causes X and is correlated with Y)

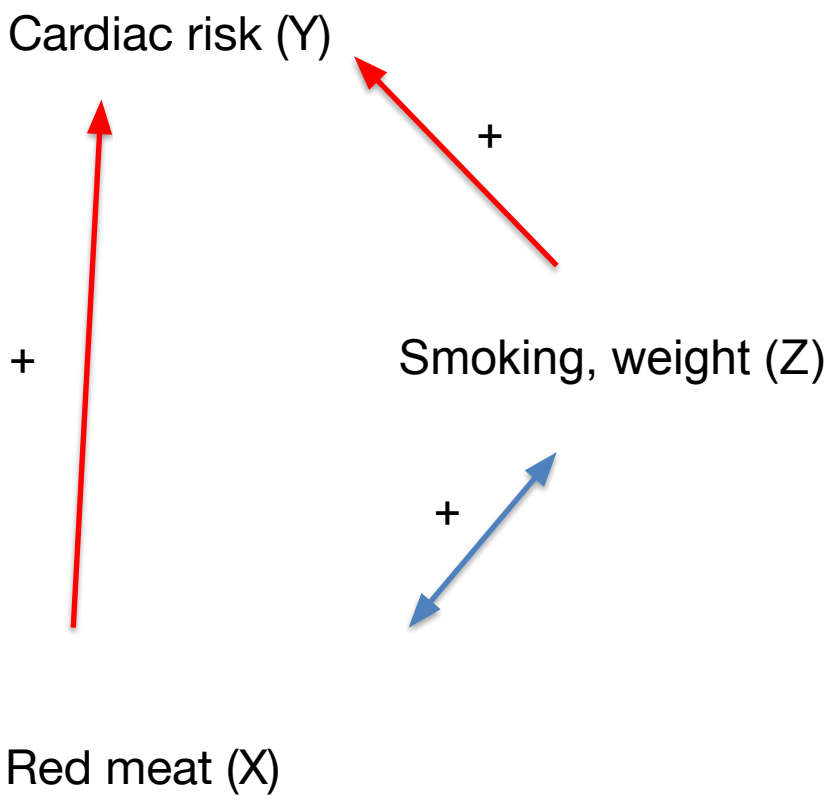
Example (1)

- The data shows positive correlation between red meat and cardiac risk
  - Y: cardiac risk
  - X: red meat
  - Z: smoking

- Possible reasons for correlation:

1. (Effect of interest) Eating red meat causes heart disease
2. (Reverse causality) Cardiac risk makes you eat more red meat (unlikely, if anything could be negative effect)
3. (Omitted variables) Red meat consumption is correlated with smoking (weight) and smoking (weight) causes heart disease (possible)

Example (1): Causal Graph



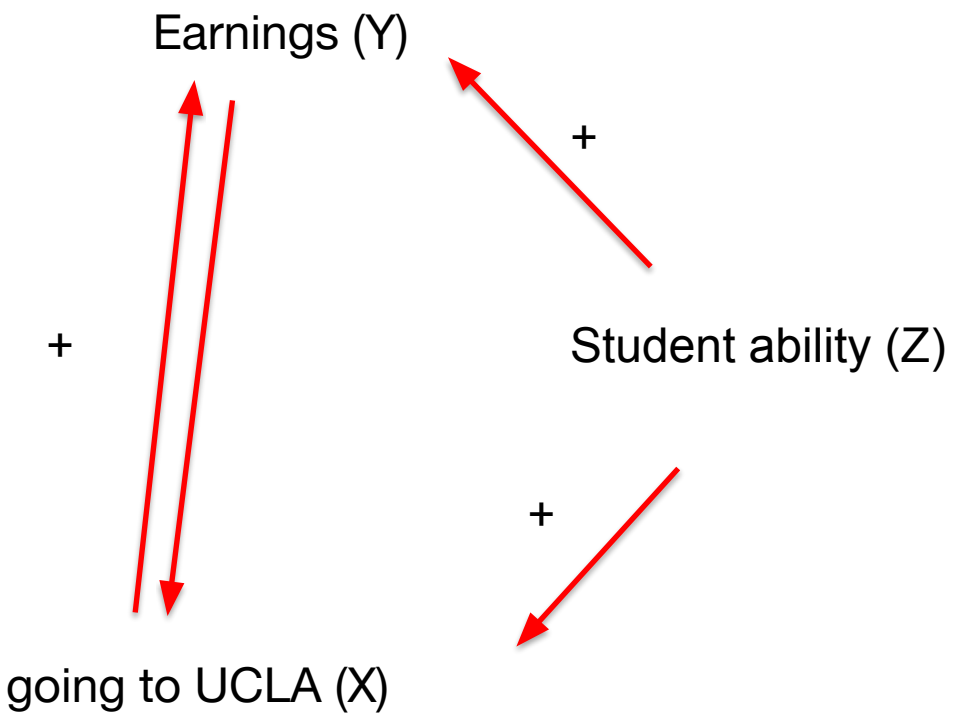
- We want to isolate the direct effect of red meat on cardiac risk
- Correlation also captures the indirect effect of smoking, weight

Example (2)

- Suppose you observe correlation in the sense that UCLA graduates earn more than USC graduates
    - Y: earnings after college
    - X: whether you studied at UCLA (X=1 for UCLA, X=0 for USC)
    - Z: student ability

- Possible reasons for correlation:
    1. (Effect of interest) A UCLA degree results in higher earnings
    2. (Reverse causality) Higher earnings affect decision to go to UCLA (possible if you choose UCLA because of expected higher earnings)
    3. (Omitted variables) Better students go to UCLA and earn more after college (maybe?)

Example (2): Causal Graph



- Note that the different mechanisms can coexist

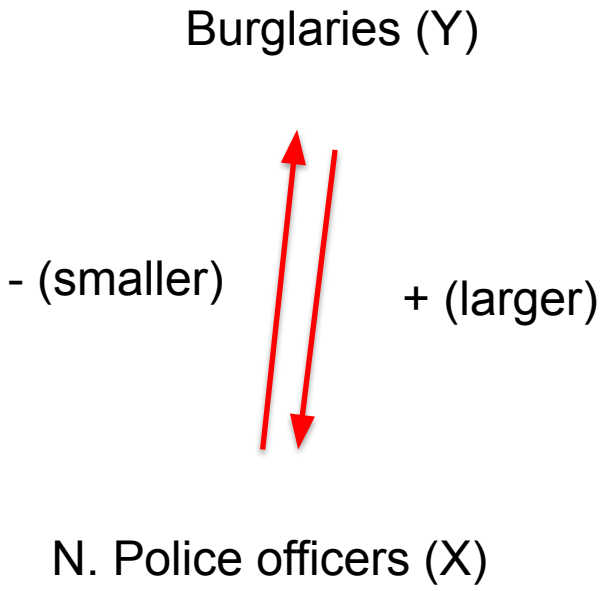
Example (3)

- Suppose that there are 2 neighborhoods: A has 10 burglaries and B has 20 burglaries, without police. The police has 10 officers. Dispatching 5 officers deters one burglary. The police dispatch all officers to neighborhood B

- What we see in the data is:
  - A: 10 burglaries, 0 police
  - B: 18 burglaries, 10 police

- We thus see a **positive correlation** between number of officers and burglaries, despite a **negative causal** effect of police on crime!

Example (3): Causal Graph



Correlation or Causation?

- **A:**
- Omitted variables?
  - exercise (must walk the dog)
- Reverse causality?
  - Not likely to get a dog if you are in very poor health

Extracting Causation from Observational Data

- Great part of this course will be about learning methods that *can help* address the endogeneity problem and disentangle causation from correlation in observational data (but you can never be sure you have causality)

- We will see that which method to use is linked to what we think are the possible reasons for endogeneity (causal channels at play). For the three previous examples we have:

    - Red meat example (1): problem is omitted variables (smoking and weight) that can be measured → we'll discuss this today in part 2

    - UCLA example (2): problem is omitted variables (ability) that are difficult to measure → we'll discuss this later in the course

Extracting Causation from Observational Data (cont.)

- Police example (3): problem was reverse causality
- **Q:** How could we estimate the causal effect of police on crime while avoiding contamination by the crime on police effect?
- **A:** The problem is that assignment of police is based on crime rates. Can we make the assignment rule not related to crime?
  - Assign police officers randomly across the city? Probably hard to find political support for such an experiment
  - Try to find historical events when the assignment process was random, e.g. due to some change in rules or other reasons not related to crime → we'll discuss this later in the course

Poll

- Say you want to estimate how house prices are affected by having a good school nearby

- You have data on each school's quality in LA and house prices in the catchment area of the school

- Say you consider only homes with a given set of characteristics (e.g., 3-bedroom) and run the regression of average prices in the catchment area for this type of home on school quality

$$House\_price = \beta_0 + \beta_1 school\_quality + e$$

slido

What is the main source of endogeneity in regressing house prices on school quality?

ⓘ Start presenting to display the poll results on this slide.

Poll answer

- Reverse causality: higher house prices means better schools (e.g. higher property taxes)

- Also omitted variable parental income/education: wealthier/educated parents pay more for homes and donate/get involved in schools

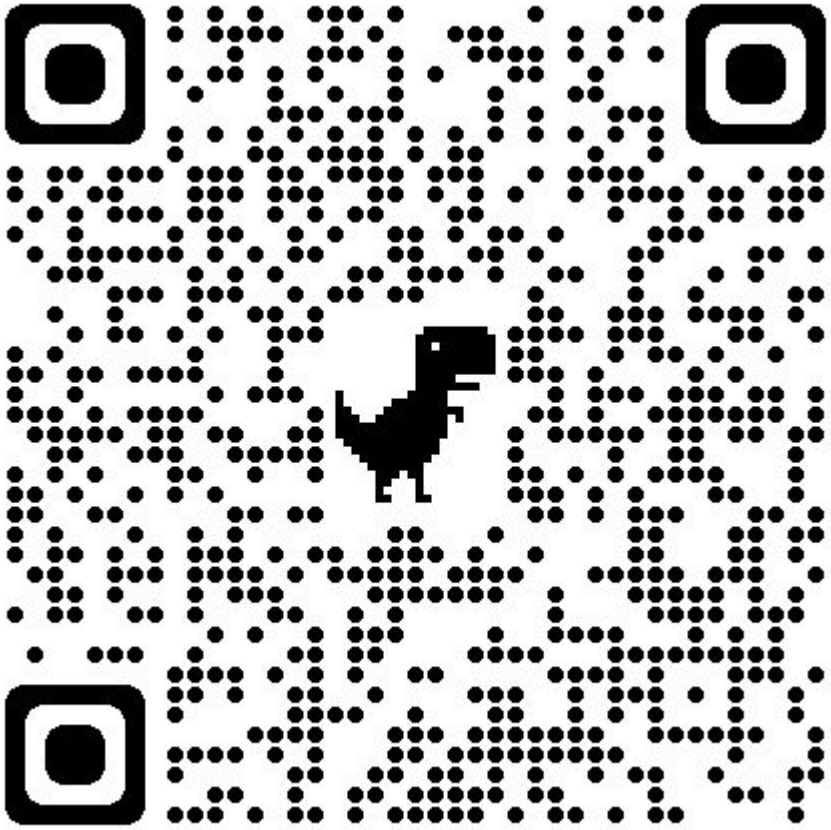
- **Q:** What about omitted house characteristics? When would they be a source of endogeneity?
- **A:** They certainly directly affect house prices, but you'd have to make the case that they are also correlated with school quality to induce endogeneity

Road map

1. Causality & correlation
2. Omitted variable bias
3. A/B testing & causality

Attendance!

Link will close soon!



Multivariate regression with controls

- Say we have observational data where we can measure omitted variables $Z_1, \ldots, Z_k$

- We can then consider a multivariate regression:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_{1i} + \cdots + \beta_{k+1} Z_{ki} + e_i$$

  - X is the variable whose causal effect we care about
  - $Z_1, \ldots, Z_k$ are control variables ("controls")

- Least-squares estimation of coefficients:

$$\min_{\beta_0,\ldots,\beta_{k+1}} \sum_i e_i^2 = \min_{\beta_0,\ldots,\beta_{k+1}} \sum_i (Y_i - (\beta_0 + \beta_1 X_i + \beta_2 Z_{1i} + \cdots + \beta_{k+1} Z_{ki} + e_i))^2$$

Red meat example (1)

- Since we can measure the omitted variable smoking in the data, run a regression of cardiac risk on red meat consumption AND smoking
- **Q:** Interpret the coefficients of the multivariate regression

```
model  = smf.ols(formula = 'cardiac_risk ~ red_meat + smoking',data=df)
result = model.fit()
print(result.summary())
```

A:

	coef	std err	t	P>|t|
Intercept	3.5018	0.080	43.742	0.000
red_meat	0.0545	0.040	1.355	0.176
smoking	0.8776	0.118	7.462	0.000

An additional serving of red meat increases the cardiac risk score by 0.05, **holding constant the amount of smoking**

→ the multivariate regression disentangles the effect of red meat from that of smoking

Omitted variable bias

- Let's assume that the true population relationship is:

$$cardiac\_risk = \beta_0 + \beta_1 read\_meat + \beta_2 smoking + e$$

With e such that $cov(red\_meat, e) = 0$

- Assume we forgot about the impact of smoking (or we lack data to measure it) and that we run the univariate regression on red meat only:

$$cardiac\_risk = \beta_0 + \beta_1 read\_meat + u$$

Q: How does smoking implicitly influence the univariate regression?

A: Smoking is now part of the error term: $u = \beta_2 smoking + e$

Omitted variable bias

- Remember the expected value of $\widehat{\beta}_1$ from the univariate regression

$$E(\widehat{\beta}_1) = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(red\_meat, cardiac\_risk)}{Var(red\_meat)}$$

- Plug into the covariance the true expression for cardiac_risk,

$$cardiac\_risk = \beta_0 + \beta_1 red\_meat + \beta_2 smoking + e$$

$$E(\widehat{\beta}_1) = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(red\_meat, \beta_0 + \beta_1 red\_meat + \beta_2 smoking + e)}{Var(red\_meat)}$$

$$= \beta_1 \frac{Cov(red\_meat, red\_meat)}{Var(red\_meat)} + \beta_2 \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

$$\Rightarrow E(\widehat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

Omitted variable bias

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

Bias

Q: What do you think is the sign of the bias? Are we over- or under-estimating the effect of red meat on cardiac risk if we omit smoking from the regression?

A: Overestimating, because probably

- Cov(red_meat,smoking)>0 (people that eat a lot of red-meat all tend to smoke more)
- β_2>0 (smoking has a positive effect on cardiac risk)

Omitted variable bias: when does it matter?

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

Q: In what case does omitting smoking **not** lead to bias?

A:

- Cov(redmeat,smoking)=0 OR
- β_2=0

Therefore: only worry about omitted variables that are correlated with the outcome variable Y (cardiac risk) AND correlated with the included variable X (red meat)

slido

The Bias due to the omitted variables "servings of vegetables per week" is:

ⓘ Start presenting to display the poll results on this slide.

Poll answer

Positive bias because cov(red_meat,veggies)<0 and effect of veggies on cardiac risk is <0 so the product is positive

Relationship to causal graph

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

- The last term is the regression coefficient from the regression:

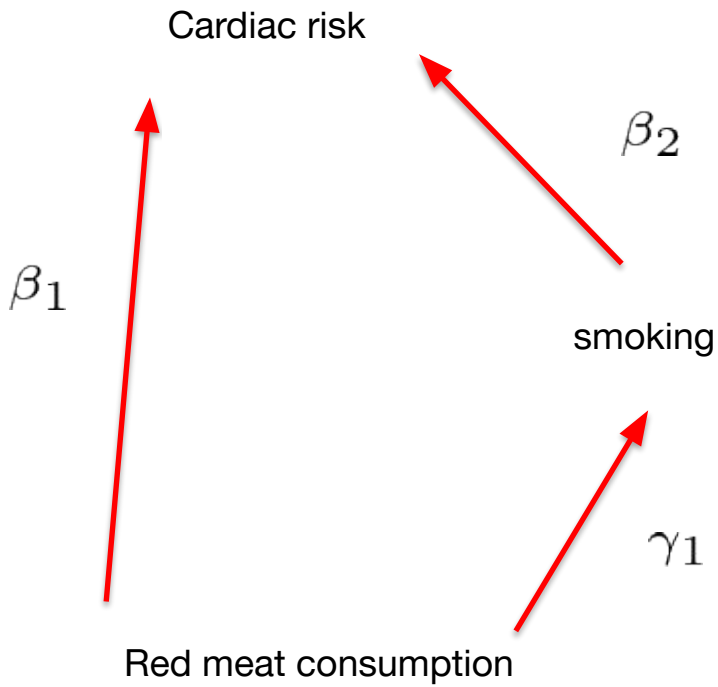
$$smoking = \gamma_0 + \gamma_1 red\_meat + \nu$$

$$\gamma_1 = \frac{Cov(red\_meat, smoking)}{Var(red\_meat)}$$

- Therefore:

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \gamma_1$$

Relationship to causal graph



- We saw on the previous slide: $E(\widehat{\beta}_1) = \beta_1 + \beta_2\gamma_1$
- Univariate regression captures both direct effect of red meat (β_1) and indirect effect due to correlation with the omitted variable smoking (β_2 γ_1)

Eliminate bias by adding the omitted variable

- Solution: include omitted variables in regression (if you can measure them)

$$cardiac\_risk = \beta_0 + \beta_1 read\_meat + \beta_2 smoking + e$$

- Now we can pin down the direct effect of red meat (provided e is uncorrelated with red meat): $E(\widehat{\beta}_1) = \beta_1$

Eliminating bias by adding the omitted variable

- Univariate regression:

	coef	std err	t	P>|t|
Intercept	3.3791	0.080	41.994	0.000
red_meat	0.3187	0.020	16.264	0.000

- Multivariate regression with control for smoking:

	coef	std err	t	P>|t|
Intercept	3.5018	0.080	43.742	0.000
red_meat	0.0545	0.040	1.355	0.176
smoking	0.8776	0.118	7.462	0.000

- An additional serving of red meat increases the cardiac risk score by 0.05, **holding constant the amount of smoking**
- Including smoking in the regression corrects for the over-estimation of the effect of red meat that was due to omitted variable bias

Poll: demand estimation

We regress sales on own price (p1) and competitor price (p2)

Univariate

```
lm(formula = Sales ~ p1, data = multi)

Coefficients:
            Estimate Std Error t value p value
(Intercept)   211.20     66.49    3.18   0.002
p1             63.71     13.04    4.89   0.000
---
Standard Error of the Regression:  223.4
Multiple R-squared:  0.196  Adjusted R-squared:  0.188
Overall F stat: 23.87 on 1 and 98 DF, pvalue= 0
```

Multivariate

```
lm(formula = Sales ~ p1 + p2, data = multi)

Coefficients:
            Estimate Std Error t value p value
(Intercept)   115.70     8.548   13.54       0
p1            -97.66     2.669  -36.60       0
p2            108.80     1.409   77.20       0
---
Standard Error of the Regression:  28.42
Multiple R-squared:  0.987  Adjusted R-squared:  0.987
Overall F stat: 3717.29 on 2 and 97 DF, pvalue= 0
```

Q:What is the correlation between p1 and p2?

slido

What is the correlation between p1 and p2?

ⓘ Start presenting to display the poll results on this slide.

Poll answer

- The coefficient for p1 goes down after including p2 in the regression, and the effect of p2 is positive
- This means that cov(p1,p2)>0 given the formula for the bias

Road map

Agenda

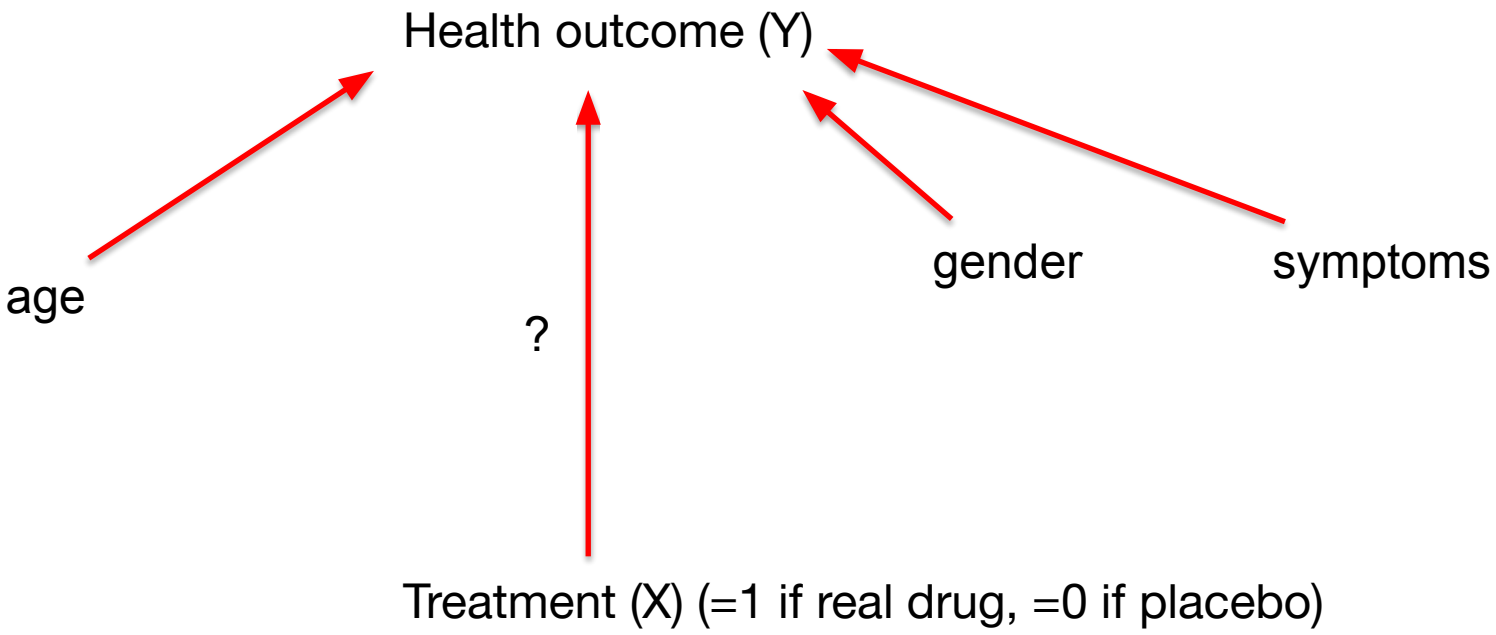
1. Causality & correlation
2. Omitted variable bias
3. A/B testing & causality

Causation in Experimental Data

- Experimental data (a.k.a. A/B testing or Randomized Controlled Trials) are the ideal data to study causality
- Think of the typical question of assessing drug effectiveness
    1. Experiment: randomly assign patients to receive a drug (treated group) or a placebo (control group)
    2. Observational data: compare all patients that took the drug in the past with patients that did not take it

- **Q:** If Y is a health outcome and X the drug, what could be omitted variables Z in these two types of data?
- **A:** In case 1. random assignment means that by construction X is NOT correlated with any Z
- In case 2. patients that took the drug could have different characteristics than patients who didn't – symptoms, age, gender, etc. that are correlated with Y

Experiment Causal Graph



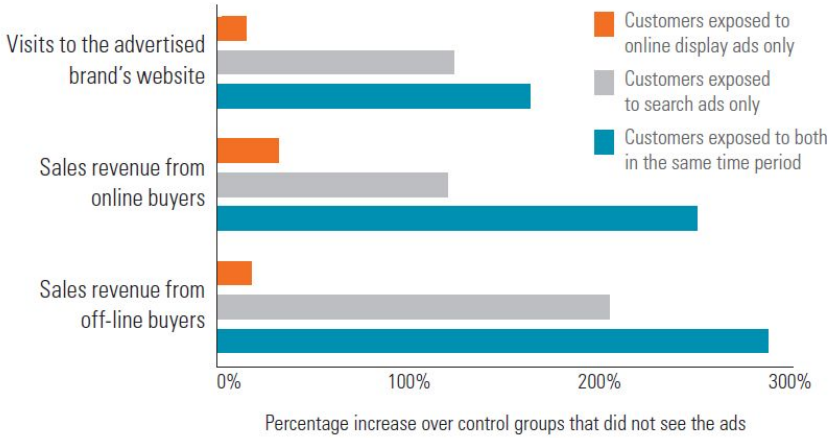
- If treatment assignment is random, then there are no arrows from treatment to the other variables

HBR study

HARVARD BUSINESS PUBLISHING Harvard Business Review

Web Ads Boost In-Store Sales, Too

Results from 18 studies in the finance, travel, telecommunications, and retail sectors collectively show that online ads have a powerful effect on off-line sales. Running search ads tends to be more effective than using display ads, and combining both types is more effective still.



Q: Would you recommend using display and search ads?

A: First you want to understand what type of studies they are talking about

HBR study

- The article describes its methodology as follows:

"Measuring the online sales impact of an online ad …is straightforward: We determine who has viewed the ad, then compare online purchases made by those who have and those who have not seen it."

- Observational study

- **Q:** What could be possible issues?

HBR Study Critique

- Problem with observational study: people who saw the search ad may be people who were searching for the product

  → targeted ad

- Concrete example:
  - You have a company that sells golf clubs and run ads on the golf channel (i.e., a targeted ad)
  - Say you have data where you can observe whether people bought or not your golf clubs and whether or not they saw the ad
  - You find that people that saw the ad were 10-times more likely to buy your golf clubs than people that did not see the ad

- Does this correctly capture the effectiveness of advertising? If not, does it over- or understate the effectiveness?

Re-frame this as omitted variable bias

- In observational data the ad effect is estimated by running the regression

$$Purchase = \beta_0 + \beta_1 Golf\,Ad + e$$

- Here both *Purchase* and *Golf Ad* variables are binary (dummy) variables

  → Linear Probability Model

- Estimate the coefficients by least squares and test the hypothesis that β_1=0

Q: Interpret the coefficients and the test

A: Intercept β_0 : Purchase probability for consumer that did not see ad

Slope β_1 : Difference in purchase probability for consumers that saw ad versus those that did not

If t-test rejects, the ad is effective in the sense that it changes the probability of purchasing

Re-frame this as omitted variable bias

$$Purchase = \beta_0 + \beta_1 GolfAd + e$$

Q: How does "watching the golf channel" influence the regression?

A: It's part of the error term, because it plausibly has a direct effect on Purchase

$$e = \beta_2 GolfChannel + u$$

… hence we have omitted variable bias since it is also correlated with GolfAd :

$$E(\widehat{\beta}_1) = \beta_1 + \frac{Cov(GolfAd, e)}{Var(GolfAd)} = \beta_1 + \beta_2 \frac{Cov(GolfAd, GolfChannel)}{Var(GolfAd)}$$

Both ≠ 0

Signing the bias

$$E(\widehat{\beta}_1) = \beta_1 + \beta_2 \frac{Cov(GolfAd, GolfChannel)}{Var(GolfAd)}$$

Q: What is the sign of the bias? Are we over- or under-stating ad effectiveness?

A: Overstate!

- Cov(GolfAd,GolfChannel)>0: People that watch the golf channel are more likely to see the ad
- β_2>0: Watching the golf channel is associated with a higher likelihood of purchasing golf clubs

Important managerial insight: overstatement is very common in measurement of ad effectiveness using observational data when ads are targeted!

Solution: experimental data (A/B test)

- Problem: in observational data people that saw the ad are different from people that did not see the ad (in a way that is related to Y)

- Solution: run an experiment (A/B test)

  - Treatment: show ads (e.g. golf club ads on the webpage of an online golf-magazine)
  - Randomization: When a potential customer arrives on the page, the computer shows the ad with a 50% probability
  - We thus have 50% of people that saw the ad (treated group) and 50% that did not see the ad (control group)

Estimating ad effect on experimental data

- Same regression as before but now based on experimental data:

$$Purchase = \beta_0 + \beta_1 GolfAd + e$$

$$E(\widehat{\beta}_1) = \beta_1 + \frac{Cov(GolfAd, e)}{Var(GolfAd)}$$

= 0 because of randomization !!!

slido

Adding the variable GolfChannel to the regression (on experimental data)

ⓘ Start presenting to display the poll results on this slide.

Poll answer

- Because of randomization *GolfAd* is uncorrelated with *GolfChannel* so the estimator of β_1 remains unbiased when adding *GolfChannel* as regressor
- Adding variables to the regression changes the estimator of the standard error so the precision changes

→ we will discuss this next week…

A final useful example – Huang et al. (2018)

- How annoying are ad interruptions in free, ad-supported websites?
- From the website perspective, interested in knowing the sensitivity of content consumption to the amount of ad interruptions (price): demand estimation

Q: How would you go about measuring it?

A: Experiment: randomly assign different amounts of interruptions to different users and measure how much content they consume (Huang et al. 2018)

Q: What regression would you run?

A:

$$ContentHours = \beta_0 + \beta_1 AdsPerHour + e$$

A final useful example – Huang et al. (2018)

- **Q:** Potential downsides of running price experiments?

- A: Customer alienation if they find out. E.g., failed experiment by Amazon in 2000-2001: “It was a mistake because it created uncertainty for customers rather than simplifying their lives” (Jeff Bezos)

- **Q:** What is the endogeneity problem when answering this questions using observational data (i.e. you only see a dataset with natural variation in interruptions and content consumption)? Possible causal channels?
- A: Opportunity cost of time is a likely omitted variable: if your cost is low, you have a lot of time to spend online and probably you care less about wasting time listening to ads

Summary

- Causality
  - In observational data, correlation does not generally equal causation
  - Experimental data (A/B tests) make causal inference possible by manipulating relevant X variable (and X alone)

- Omitted variable bias (OVB)
  - Omission of variables matters only if they correlate with both Y and X
  - If omitted variables are measurable multivariate regression gives the direct effect

- A/B tests
  - Do not suffer from OVB because treatment variable X not correlated with **any** other variables
  - This is by design because treatment is randomly assigned