Applied Data Analysis (CS401)

Lecture 6
Observational studies
21 Oct 2020

Robert West

Announcements

- Projects milestone P1 due this Fri 23 Oct 23:59
- Projects milestone P2 released the same day
  - Your first data analysis replication, yay!
  - Due 2 weeks later (Fri 6 Nov)
- Friday's lab session:
  - More + exercises on observational studies

Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2020-lec6-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

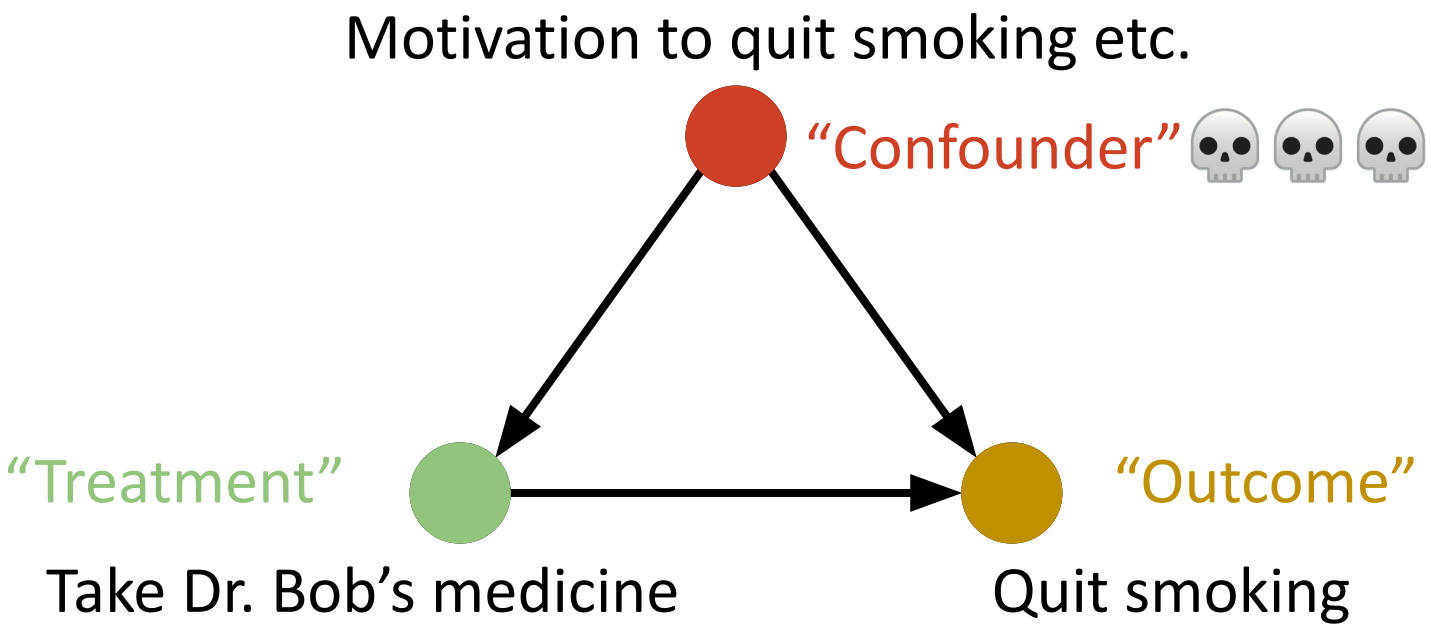
Dr. Bob's smoking cure

- I claim to have developed a medicine that helps you quit smoking
- I ask all smokers: "Do you want to try my medicine?"
- Smokers = treated smokers $\cup$ untreated ("control") smokers
- Fraction of successful quitters is higher in the treated group
- I conclude: "My medicine helps you quit smoking! Buy it!"
- Do you believe me?

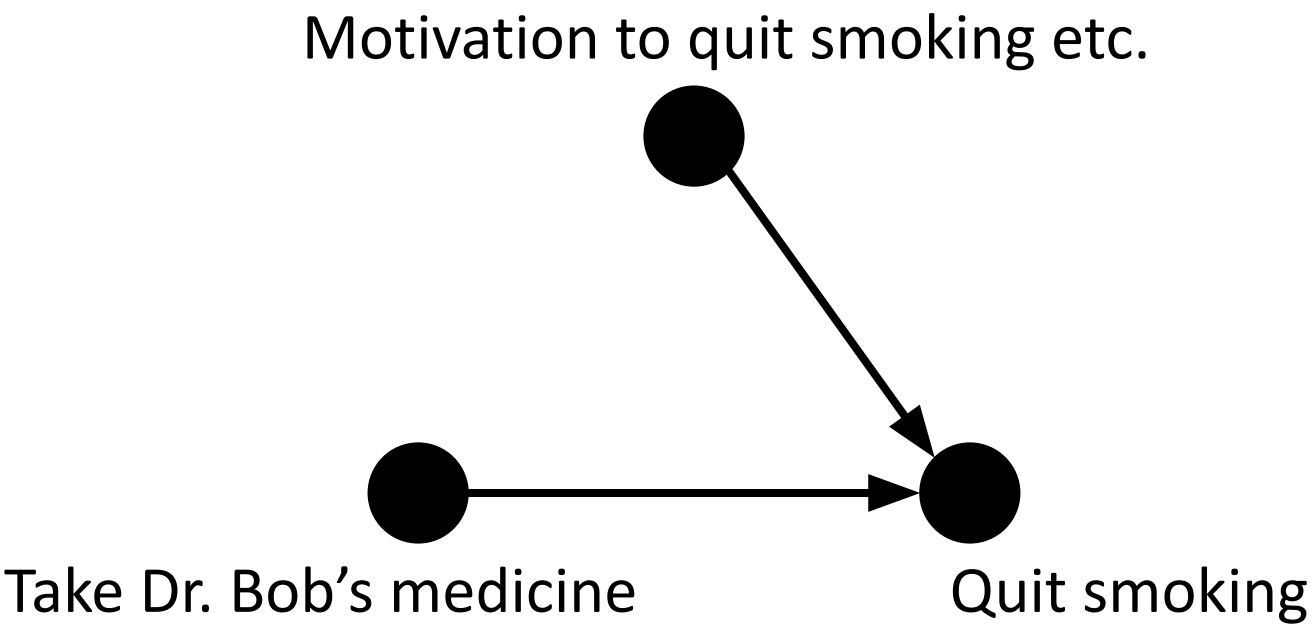
Goals of this lecture

- Clarify difference between **experimental** and **observational** studies
- Highlight **pitfalls** of observational studies
- Give you **tools** for drawing valid conclusions from "found data" (very useful for project!)
- Motivate you to **read** Rosenbaum's great book "Design of Observational Studies" (in particular Chapters 1, 2, and 3; or this book) and Pearl's eye-opening "Book of Why"

Dr. Bob's "experiment" as a causal diagram



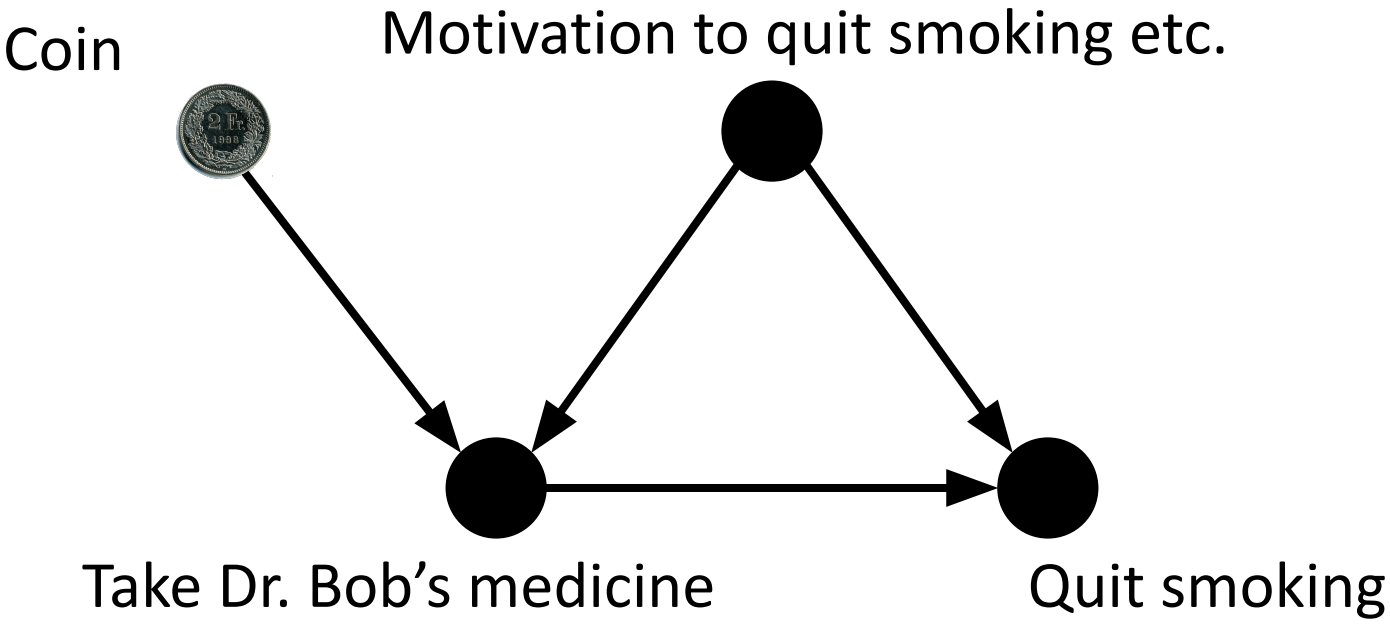
Ideal setting as a causal diagram



Randomized controlled experiments

- Two experimental conditions:
  - Treatment (e.g., medicine)
  - Control (e.g., placebo [fun fact])
- Assignment of participants to conditions is random
  - Probability of receiving treatment same for everyone
- Treatment and control groups are indistinguishable
  - E.g., determination to quit smoking is not systematically higher in the treated group

Randomized controlled experiment as a causal diagram



Limits of randomization

- Do seat belts save lives?
- Experiment:
  - Flip coin at birth to assign to treatment (always wear seat belt for entire life) or control (never wear seat belt)
  - Measure fraction of traffic deaths in each group
- Randomized experiments aren't always feasible
  - Unethical (see above), expensive, fundamentally impossible (e.g., do earthquakes decrease life spans?)
  - Most modern "big data" is "found data"
- Sometimes, observational studies are even better suited

Alternative: observational studies

- Fundamentally different from experiment:
  - Researcher can’t control who goes to which condition
  - Researcher is merely an observer, not a tinkerer
  - Much less problematic w.r.t. ethics, price, feasibility
  - Much more problematic w.r.t. validity of conclusions
- All advantages of randomized experiment are gone
  - Subjects self-select to be treated
  - Treatment assignment and response may be caused by same hidden correlate (a.k.a. confounders; e.g., resolve to quit smoking)

Example: seat belts revisited

- Recall: experiment infeasible because unethical
- Observational study:
  - Dataset: all traffic accidents in a given time span
  - Two treatment conditions:
    - Treated: seat-belt wearers
    - Control: non-seat-belt wearers
  - Compare fraction dead in treated vs. control
- What problems do you see?

- Two treatment conditions:
  - Treated: seat-belt wearers
  - Control: non-seat-belt wearers
- Compare fraction dead in treated vs. control

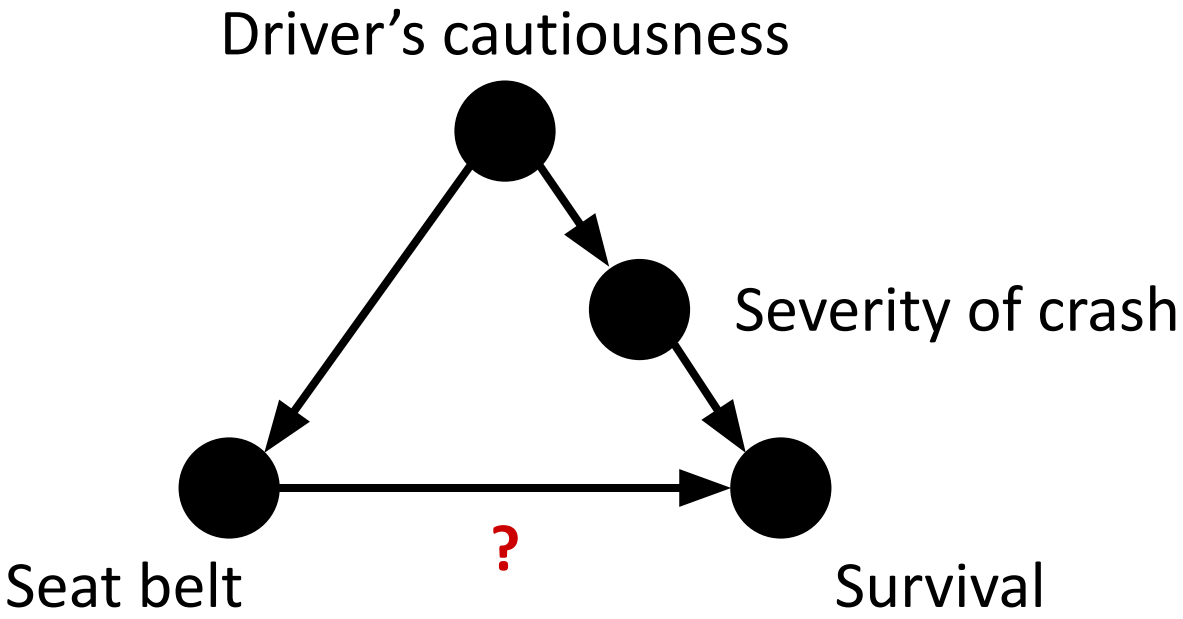
- What problems do you see?

CHAT ROULETTE!

Think for 1 minute!

- Then: chat with a fellow student for 3 minutes
  - Rolex Forum: Talk to neighbor (priority: left, right)
  - Zoom: You'll be randomized into small **"breakout rooms"**

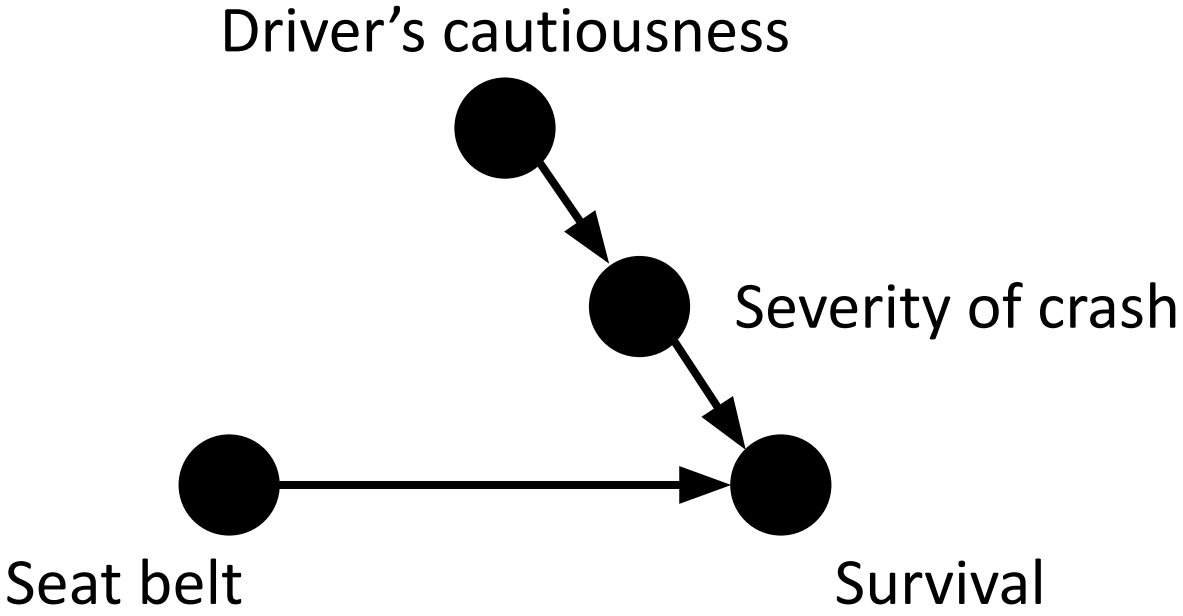
As a causal diagram



A matched observational study

- Consider only particular subset of accident cars:
  - 2 people in car: driver + passenger
  - Exactly one of them died in accident
  - Exactly one of them wearing seat belt at time of accident (i.e., 1 treated + 1 control per car)
- As before: compare fraction dead in treated vs. control
- New: everything else is controlled for, incl. type of car, speed, severity of accident
- Fundamental concept: **matching**

As a causal diagram



Settling the seat-belt question

	Driver Passenger	Not Belted Belted	Belted Not Belted
Driver Died	Passenger Survived	189	153
Driver Survived	Passenger Died	111	363

Natural experiments

- Not researcher, but nature, “flips a coin” to decide treatment assignment
- Rosenbaum: “When investigators are especially proud, having found unusual circumstances in which treatment assignment, though not random, seems unusually haphazard, they may speak of a ‘natural experiment.’”
- E.g., seat-belt study: ceteris paribus, who wears the belt is (nearly) haphazard

Nature didn’t flip a coin for me – should I just go home and weep?

- No! You can still get good mileage if you’re smart about it
- Fundamental concept: matching
- Idea: Pair up 2 “similar” people: 1 treated, 1 control
- Ideal (rather: Utopian): “similar” := “identical”
- Also sufficient (phew!): “similar” := equal probability to receive treatment (given the state of the world before the study)

Time for some notation

$$\pi_\ell = \Pr\left(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell\right)$$

ℓ : a subject participating in the study

π_ℓ : probability of being treated, *given full knowledge of the world*

Z_ℓ : treatment assignment (1 := treated; 0 := control)

$r_{T\ell}$: response if subject is treated (observed iff $Z = 1$)

$r_{C\ell}$: response if subject is control (observed iff $Z = 0$)

$\mathbf{x}_\ell$: observed covariates (a.k.a. features)

u_ℓ : unobserved covariates

The ideal matching

- Recall: we match 1 treated with 1 control subject
- Ideal matching: equal probability to be treated:
  $\pi_k = \pi_\ell$ for all matched pairs (k, ℓ)
- Why is this ideal? Because it entails that each individual is equally likely to be the treated one in the pair

$$\Pr(Z_k = 1, Z_\ell = 0 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell, Z_k + Z_\ell = 1)$$

$$= \frac{\Pr(Z_k = 1, Z_\ell = 0 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)}{\Pr(Z_k + Z_\ell = 1 \,|\, r_{Tk}, r_{Ck}, \mathbf{x}_k, u_k, r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell)}$$

$$= \frac{\pi_\ell^{1+0}\,(1-\pi_\ell)^{(1-1)+(1-0)}}{\pi_\ell^{1+0}\,(1-\pi_\ell)^{(1-1)+(1-0)} + \pi_\ell^{0+1}\,(1-\pi_\ell)^{(1-0)+(1-1)}}$$

$$= \frac{\pi_\ell\,(1-\pi_\ell)}{\pi_\ell\,(1-\pi_\ell) + \pi_\ell\,(1-\pi_\ell)} = \frac{1}{2}$$

Ok, so are we done?

- Problem: You generally don't know the probabilities to treat:

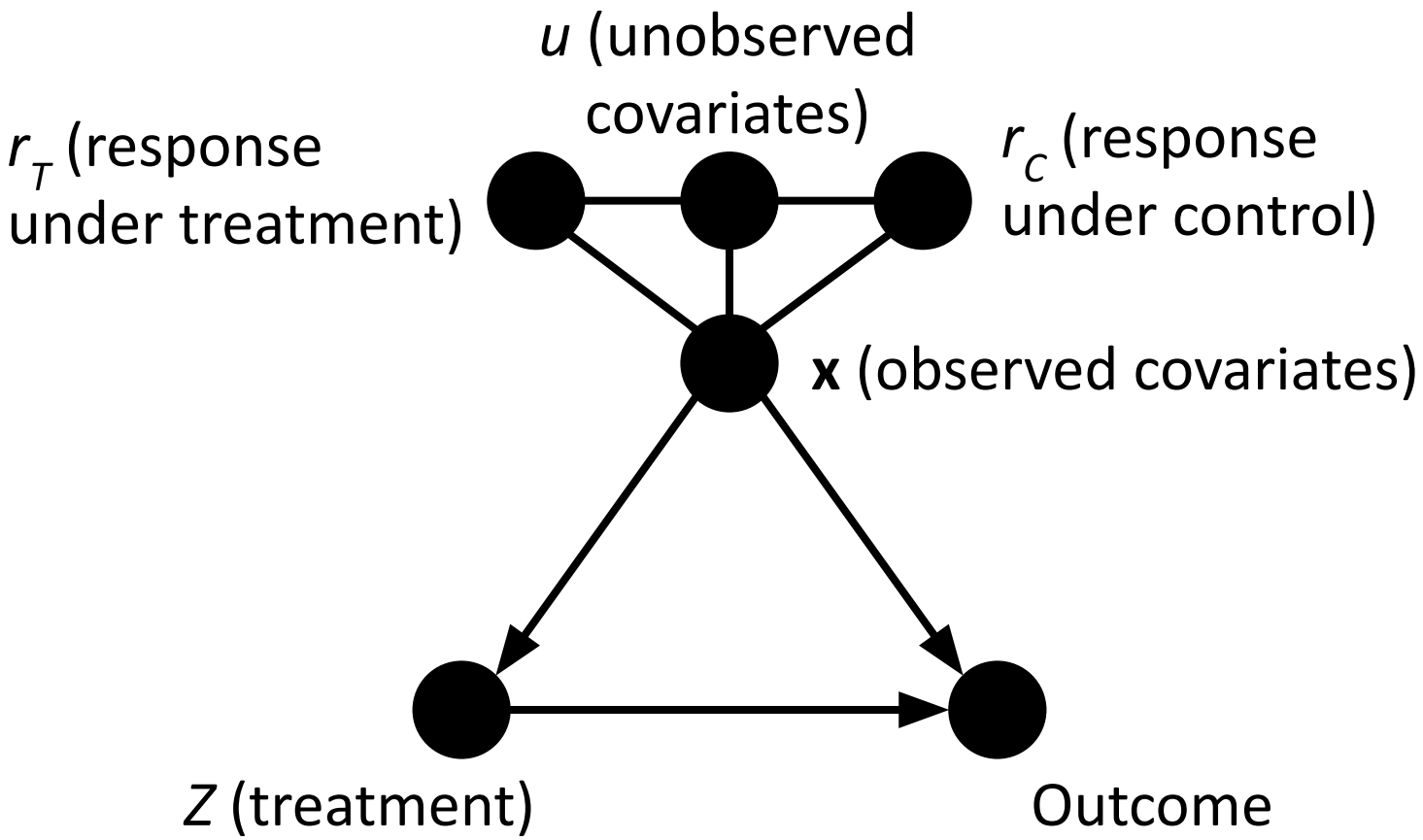
$$\pi_\ell = \Pr(Z_\ell = 1 \mid \boxed{r_{T\ell}}, \boxed{r_{C\ell}}, \mathbf{x}_\ell, \boxed{u_\ell})$$

- A **naive model:** "People who look comparable are comparable", or "Only observed variables determine treatment assignment":

$$\pi_\ell = \Pr(Z_\ell = 1 \mid \cancel{r_{T\ell}}, \cancel{r_{C\ell}}, \mathbf{x}_\ell, \cancel{u_\ell}) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell)$$

i.e., $Z \perp\!\!\!\perp (r_T, r_C, u) \mid \mathbf{x}$.

Naive model as a causal diagram



If the naive model were true…

- … you could “simulate” a randomized experiment:
  - Simply match subjects with identical observed variables **x**
  - Subjects in a pair have the same probability to treat
  - So who gets treated is up to chance, as in experiment
  - Analysis: compare outcome for treated to outcome of control (e.g., mean difference treated-minus-control)

Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
   - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
   - Solution: propensity scores
2. The naive model is naive and rarely true
   - Solution: sensitivity analysis

Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
   - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
   - Solution: propensity scores
2. The naive model is naive and rarely true
   - Solution: sensitivity analysis

Propensity scores

- Matching on observed covariates **x** is rarely feasible
- Idea: reduce the information to a single number
- Definition: propensity score $\boxed{e(\mathbf{x}) = \Pr(Z = 1 \mid \mathbf{x})}$
- Defined even when naive model not true
- But if naive model is true, it equals the probability to treat

$$\pi_\ell = \Pr(Z_\ell = 1 \mid r_{T\ell}, r_{C\ell}, \mathbf{x}_\ell, u_\ell) = \Pr(Z_\ell = 1 \mid \mathbf{x}_\ell) = e(\mathbf{x}_\ell)$$

- Typically computed via logistic regression
  - Features: **x**; label: *Z*

Balancing property of propensity score

- Fact: all subjects (treated and control) with equal propensity score have equal distribution of observed covariates **x**:

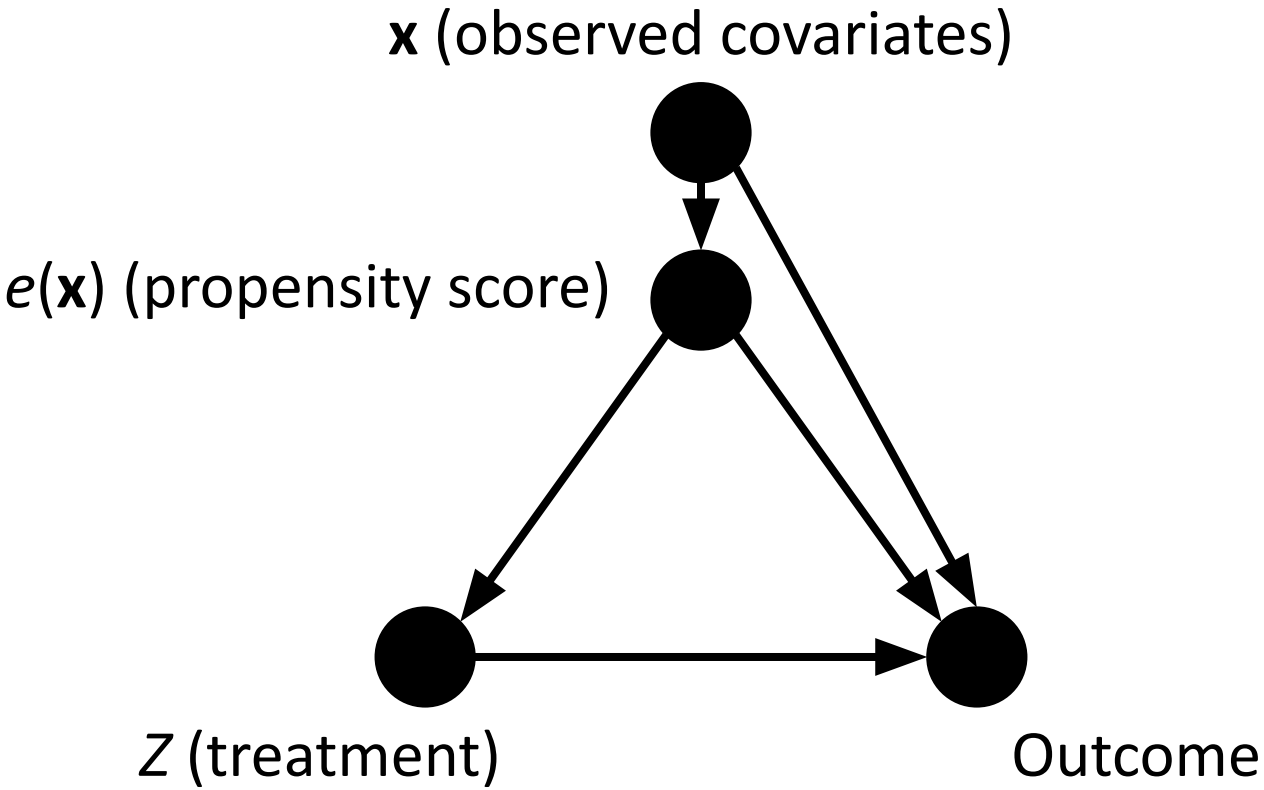
$$\Pr\{\mathbf{x} \mid Z = 1, e(\mathbf{x})\} = \Pr\{\mathbf{x} \mid Z = 0, e(\mathbf{x})\}$$

or equivalently

$$Z \perp\!\!\!\perp \mathbf{x} \mid e(\mathbf{x}),$$

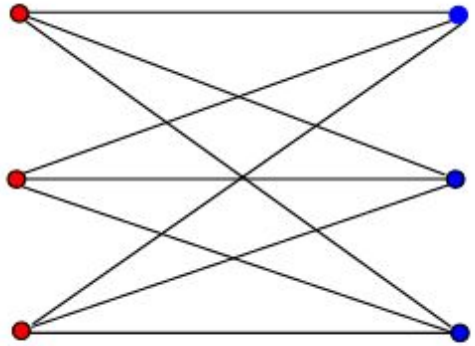
- Subjects in a matched pair might not have equal **x**, but treated and control groups will have similar distributions of **x**
- That is, matching on $e(\mathbf{x})$ is just as good as matching on **x**

Balancing as a causal diagram



Matching

- Unlikely that 2 subjects have identical propensity scores $e(\mathbf{x})$
- $\longrightarrow$ Matching (as you know it from your algorithms class)
- Bipartite graph: each subject connected to all other subjects
- Edge weights: absolute (or squared) difference of propensity scores
- Find minimum matching,
  e.g., via Hungarian algorithm



Two problems

1. Even if naive model were true: matching on **x** exactly may not be possible
   - E.g., if **x** contains 20 binary features: 1 million possible instantiations of **x**, so likely no match
   - Solution: propensity scores
2. The naive model is naive and rarely true
   - Solution: sensitivity analysis

The naive model is easily attacked

Rosenbaum:

> It is common for a critic to argue that, in a particular study, the naïve model may be false. Indeed, it may be false. Typically, the critic accepts that the investigators matched for the observed covariates, **x**, so treated and control subjects are seen to be comparable in terms of **x**, but the critic points out that the investigators did not measure a specific covariate u, did not match for u, and so are in no position to assert that treated and control groups are comparable in terms of u. This criticism could be dismissed in a randomized experiment — randomization does tend to balance unobserved covariates — but the criticism cannot be dismissed in an observational study. This difference in the unobserved covariate u, the critic continues, is the real reason outcomes differ in the treated and control groups: it is not an effect caused by the treatment, but rather a failure on the part of the investigators to measure and control imbalances in u. Although not strictly necessary, the critic is usually aided by an air of superiority: "This would never happen in my laboratory."

The sensitivity analysis model

- Idea: Quantify the degree to which the naive model is wrong
- More concretely, model assumes that treatment odds of two identically-looking subjects (i.e., identical observed covariates **x**) differ by a bounded factor Γ
- Then reasoning: "To change the conclusions of my study, two identically-looking people (1 treated, 1 control) would have to have hugely different treatment odds (i.e., huge Γ). Common sense (or domain knowledge) suggests that this is not the case, so my conclusions stand."

The sensitivity analysis model

$$\frac{1}{\Gamma} \leq \frac{\pi_k/(1-\pi_k)}{\pi_\ell/(1-\pi_\ell)} \leq \Gamma \text{ whenever } \mathbf{x}_k = \mathbf{x}_\ell . \qquad \Gamma \geq 1.$$

- Bounded odds ratio
- Odds isomorphic to probabilities
  - e.g., prob 2/3 = odds 2/1; prob 1/2 = odds 1/1
- Sensitivity $\Gamma = 1 \rightarrow$ naive model is true
- Sensitivity $\Gamma = 2 \rightarrow$ subject with same observed covariates **x** up to twice as likely to be the one to receive treatment
- Sensitivity $\Gamma = \infty \rightarrow$ void statement

Example: smoking and lung cancer

- Under naive model: matching on observed covariates gives a very small *p*-value for the null hypothesis, which states that smoking does not increase lung cancer risk (using an appropriate hypothesis test)
- Tobacco lobby: “The naive model isn’t true! There may be hidden (e.g., genetic) correlates that increase both the probability to enjoy smoking and the probability of lung cancer. They, not smoking, cause cancer!”

Example: smoking and lung cancer

- Under sensitivity analysis model, increasing sensitivity Γ increases the p-value for null hypothesis
- Anti-tobacco lobby: "But making $p > 0.05$ would require $\Gamma > 6$; i.e., the odds of being a smoker would need to be six times higher for one of two people with the exact same observed features (age, gender, education, income, ...). It's unlikely that any unobserved covariate would have such a large effect on smoking habits. So smoking causes cancer!"

Test the West!
West
West
West
Super Geschmack. Super Preis.
Der Bundesgesundheitsminister: Rauchen gefährdet Ihre Gesundheit. Der Rauch einer Zigarette dieser Marke enthält: West 1,0 mg Nikotin und 15 mg Kondensat (Teer).
West «100» 1,2 mg N. und 15 mg K. (Durchschnittswerte nach DIN).

„... und dann fand ich ein Angebot nach meinem Geschmack.“
9,90 €
34 STK
West
ORIGINAL
West
SILVER
Rauchen ist tödlich
Made for good times
LIDL
TEXAS

Two parts: mechanical vs. scientific

- Mechanical part:
  - Create pairs (1 treated + 1 control) with similar observed covariates (using exact or propensity-score matching)
- Scientific (i.e., fun) part:
  - Mitigate concerns that your findings might be caused by unobserved covariates, rather than treatment (e.g., using sensitivity analysis, ad-hoc arguments, natural experiments)

Summary

- Holy grail: randomized experiment
- When experiment not possible: observational study
- Crucial problem: treatment assignment not random (biases!)
- Semi-holy grail: natural experiment
- Matched studies: pair up treated/control based on observed covariates
- Problem: still, treatment assignment not random (biases via unobserved covariates)
- Solution: sensitivity analysis
- Keep this lecture (more here) in mind for your projects!

Blog post

"When Sheep Shop"

? How prevalent are herding effects on product review websites, and how bad are they?

Blog post

“When Sheep Shop”

? How prevalent are herding effects on product review websites, and how bad are they?

Beeradvocate®

First rating: **4.4**/5.0

Beeradvocate®

First rating: **2.7**/5.0

Blog post

“When Sheep Shop”

? How prevalent are herding effects on product review websites, and how bad are they?

Beeradvocate®

First rating: **4.4**/5.0

RateBeer

First rating: **2.7**/5.0

Blog post

“When Sheep Shop”

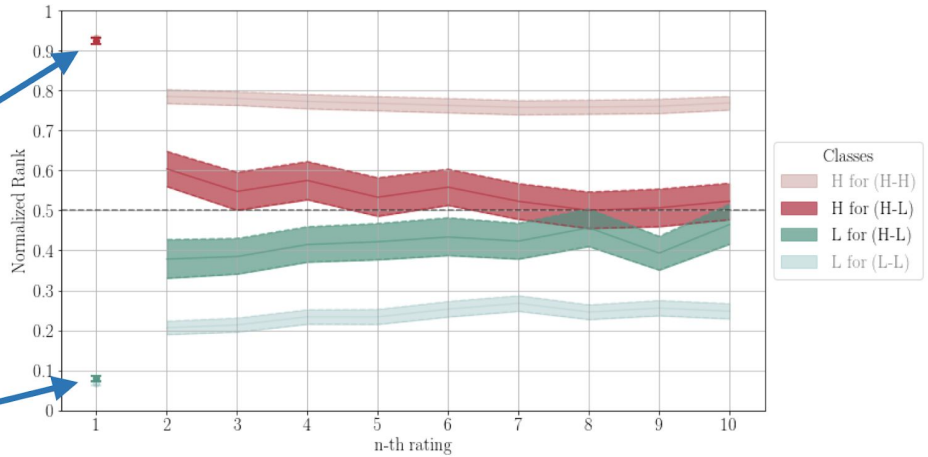
? How prevalent are herding effects on product review websites, and how bad are they?

Beeradvocate

First rating: **4.4**/5.0

RateBeer

First rating: **2.7**/5.0



Feedback

Give us feedback on this lecture here:
https://go.epfl.ch/ada2020-lec6-feedback

- What did you (not) like about this lecture?
- What was (not) well explained?
- On what would you like more (fewer) details?
- …

Credits

- Much of the material is based on Paul Rosenbaum's amazing book "Design of Observational Studies", available for free here