CSC 2541: Machine Learning for Healthcare

Lecture 3: Causal inference with observational data

Professor Marzyeh Ghassemi, PhD
University of Toronto, CS/Med
Vector Institute

VECTOR INSTITUTE | INSTITUT VECTEUR

UNIVERSITY OF TORONTO

Schedule

Jan 10, 2019, Lecture 1: Why is healthcare unique?

Jan 17, 2019, Lecture 2: Supervised Learning for Classification, Risk Scores and Survival

Jan 24, 2019, Lecture 3: Causal inference with observational data

Jan 31, 2019, Lecture 4: Fairness, Ethics, and Healthcare

Feb 7, 2019, Lecture 5: Clinical Time Series Modelling (Homework 1 due at 11:59 PM on MarkUs)

Feb 14, 2019, Lecture 6: Clinical Imaging (Project proposals due at 5PM on MarkUs)

Feb 21, 2019, Lecture 7: Clinical NLP and Audio

Feb 28, 2019, Lecture 8: Clinical Reinforcement Learning

Mar 7, 2019, Lecture 9: Missingness and Representations

Mar 14, 2019, Lecture 10: Generalization and transfer learning

Mar 21, 2019, Lecture 11: Interpretability / Humans-In-The-Loop / Policies and Politics

Mar 28, 2019, Course Presentations

April 4, 2019, Course Presentations (Project report due 11:59PM)

Outline

Slides today are courtesy of Shalmali Joshi and Uri Shalit!

1. What is confounding?
2. Why causal reasoning?
3. Potential Outcomes and Propensity Scoring
4. Pearlean Causal Graphs Framework
5. Project ideas

Outline

Slides today are courtesy of Shalmali Joshi and Uri Shalit!

1. **What is confounding?**
2. Why causal reasoning?
3. Potential Outcomes and Propensity Scoring
4. Pearlean Causal Graphs Framework
5. Project ideas

Motivational Questions

- Find which medication A/B is best for diabetics?
- Should I deploy this new feature in company’s product?
- Would this person be rejected for the job had their name been different?

Bring in the Machine Learning Hammer

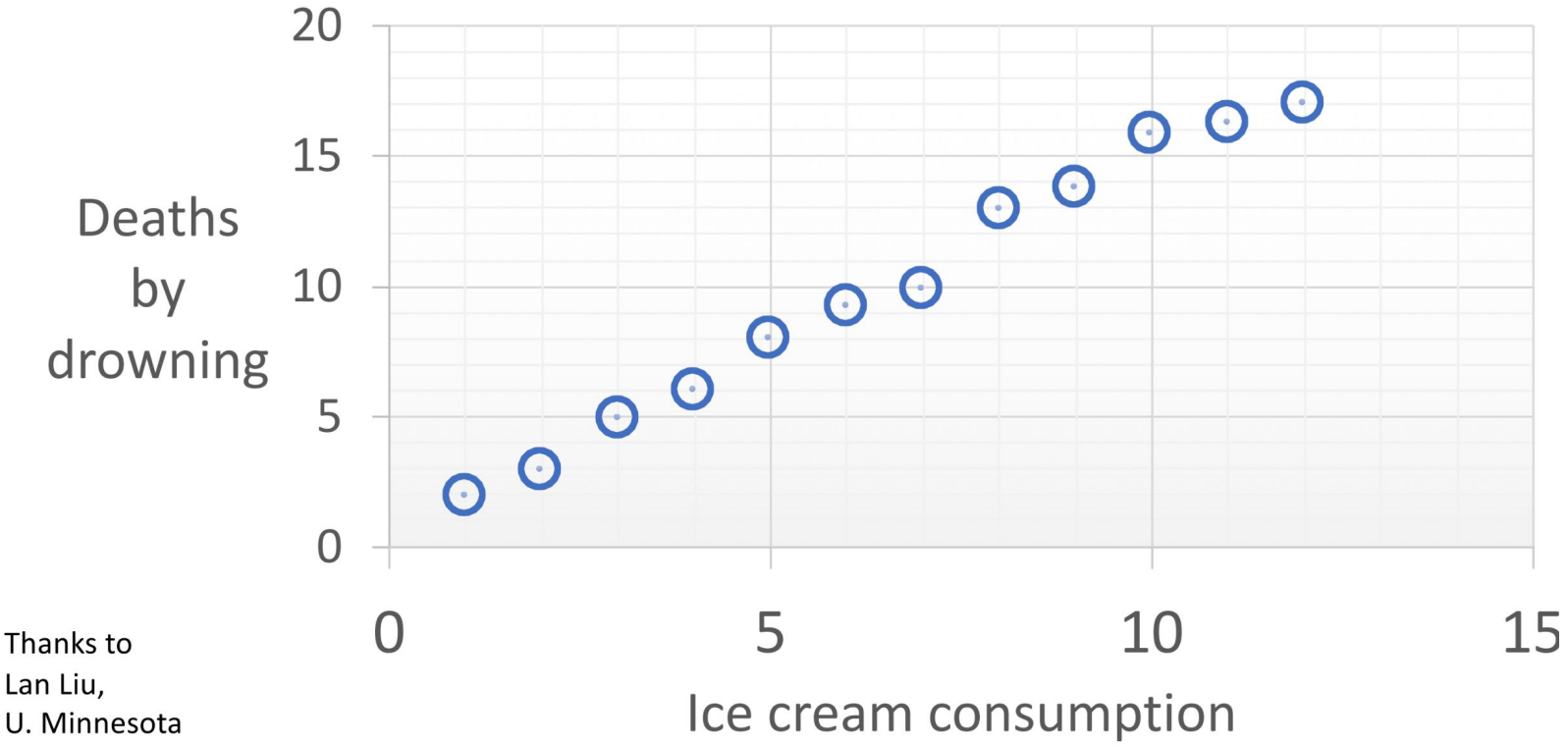
- Supervised Classification only learns “associations” p(y|x)

  X = [lab_tests, diagnoses, medications]

  y = [severely_diabetic]

- Mostly just correlations

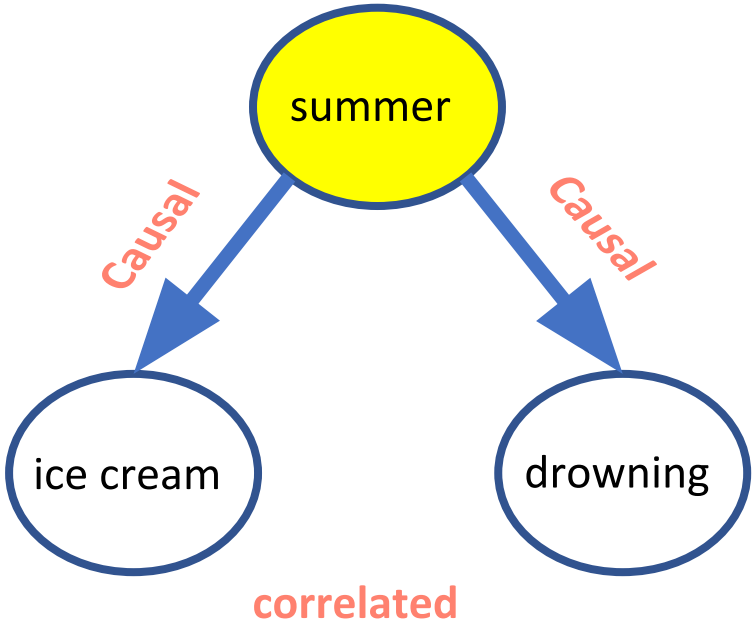
Can you spot the confounding?



Thanks to
Lan Liu,
U. Minnesota

Consider the Optics?

- Does eating ice-cream cause death by drowning?
- Is something else causing both these phenomena?
- Could we realistically have some randomly chosen humans eat lots of ice-cream and see if what happens?
- In a healthcare setting, one cannot risk death because of the treatment!



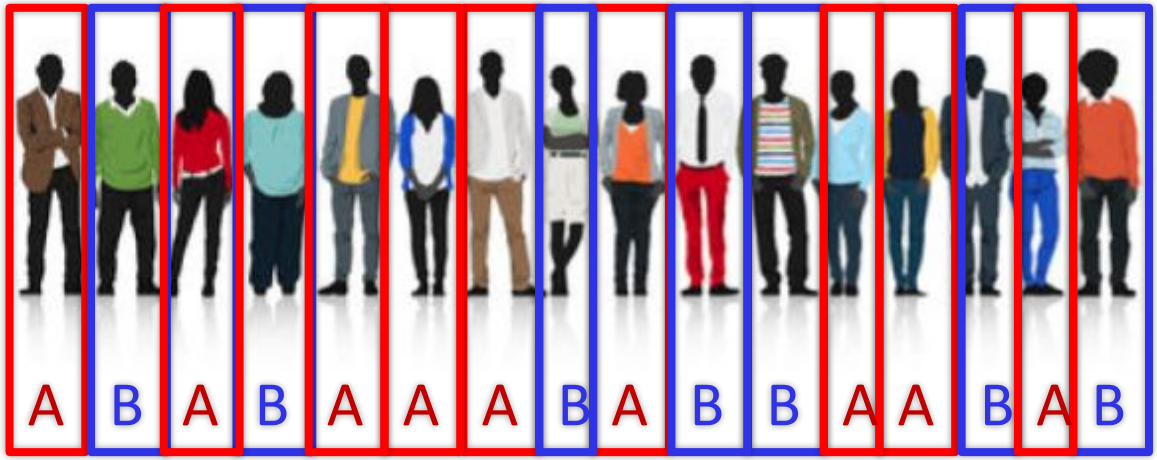
Confounding!

Randomized Controlled Trials Vs. Observational Data

treatment

A **or** B

Randomized Controlled Trials



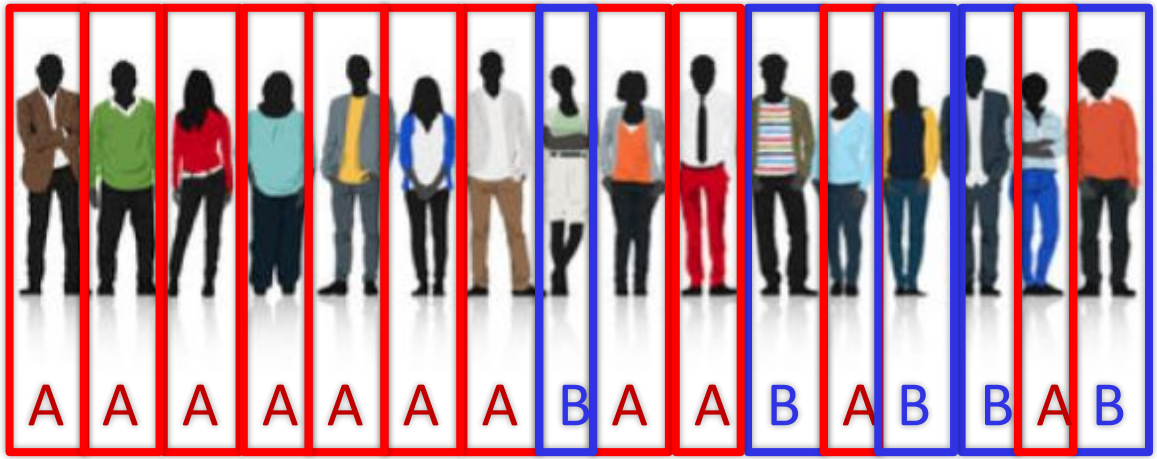
Socio-economic class

Poor → *Wealthy*

treatment

A **or** B

More Common: Observational Setting



Poor → *Socio-economic class* → *Wealthy*

treatment

A **or** B

Clinical Setting

- RCTs are also known as “clinical trials”
  - Tens of thousands every year, costing tens of billions of dollars
  - Every new medication must pass several stages of RCTs before approval for human use

- Observational study
  - Use existing data, tracking people’s medications and blood sugar
  - Problem: the space of possible confounders

Outline

Slides today are courtesy of Shalmali Joshi and Uri Shalit!

1. What is confounding?
2. **Why causal reasoning?**
3. Potential Outcomes and Propensity Scoring
4. Pearlean Causal Graphs Framework
5. Project ideas

Supervised Learning Isn’t Enough

- This is not a classic supervised learning problem; our model was optimized to predict outcome, not to differentiate the influence of A vs. B
- What if our high-dimensional model threw away the feature of medication A/B?
- Hidden confounding: Maybe using B is worse than A, but rich patients usually take B and richer people also have better health outcomes.
- If we don’t know whether a patient is rich or not, we might conclude B is better

Causal Hierarchy (not captured by mere associations)

- Observational Questions: “What if we see A”
- Action Questions: “What if we do A?”
- Counterfactuals Questions: “What if we did things differently?”
- Options: “With what probability?”

Two foundational ways to think of Causality

- Potential Outcomes (Rubin, Neyman)
- Causal Graphical Models (Judea Pearl)

Either framework requires manipulating reality

Outline

Slides today are courtesy of Shalmali Joshi and Uri Shalit!

1. What is confounding?
2. Why causal reasoning?
3. **Potential Outcomes and Propensity Scoring**
4. Pearlean Causal Graphs Framework
5. Project ideas

Potential Outcomes

- Unit: a person, a bacteria, a company, a school, a website, a family, a piece of metal, …
- Treatments / actions / interventions (A/B)
- Potential outcomes

  Y1: the unit's outcome had they been subjected to treatment t=1

  Y0: the unit's outcome had they been subjected to treatment t=0. If number of treatments is T, we have T potential outcomes (T possibly infinite)
- In observations, a single unit gets one of the T treatments

Inferring under this framework requires assumptions

SUTVA: Stable Unit Treatment Value Assumption

- The potential outcomes for any unit do not vary with the treatments assigned to other units

  Failure example: vaccination, network effects

- For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes

  Failure example: some people get out-of-date medication

- Consistency: $p(Yt=y|X=x, T=t) = p(Y = y| X=x, T=t)$

Potential Outcomes Formalized

- Sample of units $i = 1, \dots, n$
- Each has potential outcomes $(Y_0^1, Y_1^1), \dots, (Y_0^n, Y_1^n)$
- Individual Treatment Effect for unit i:

$$ITE_i \equiv Y_1^i - Y_0^i$$

- Average Treatment Effect over the sample

$$ATE_{finite} \equiv \frac{1}{n}\sum_{i=1}^{n} Y_1^i - Y_0^i$$

- Usually: assume some joint distribution $p(Y_0, Y_1)$

$$ATE \equiv \mathbb{E}[Y_1 - Y_0]$$

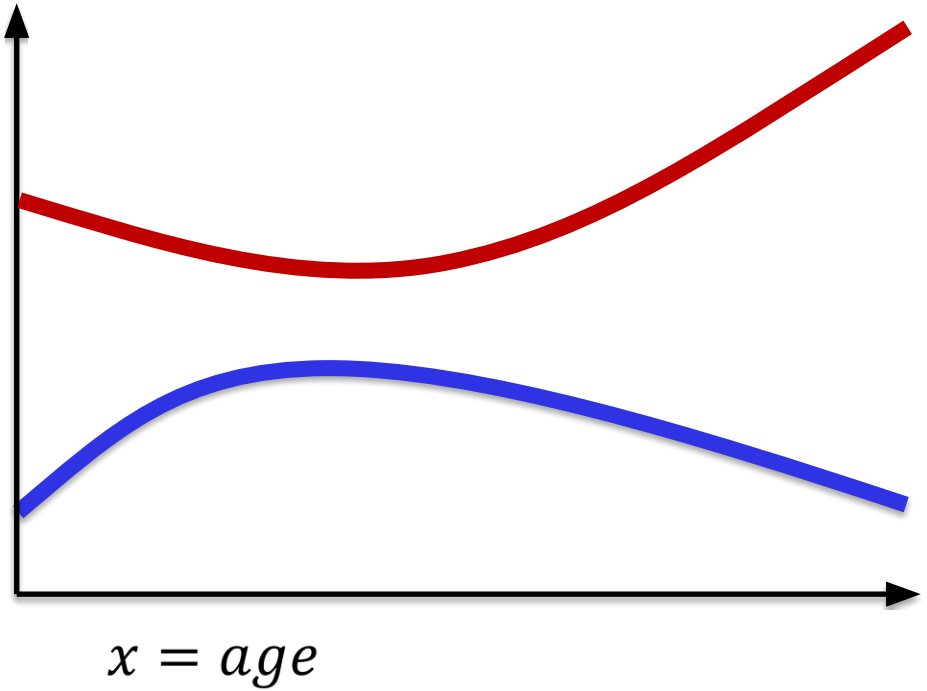
- Define average over which population ("diabetics living in Israel over age 65")

Blood Pressure and Age

$y = blood\_pres.$

$Y_1(x)$

$Y_0(x)$

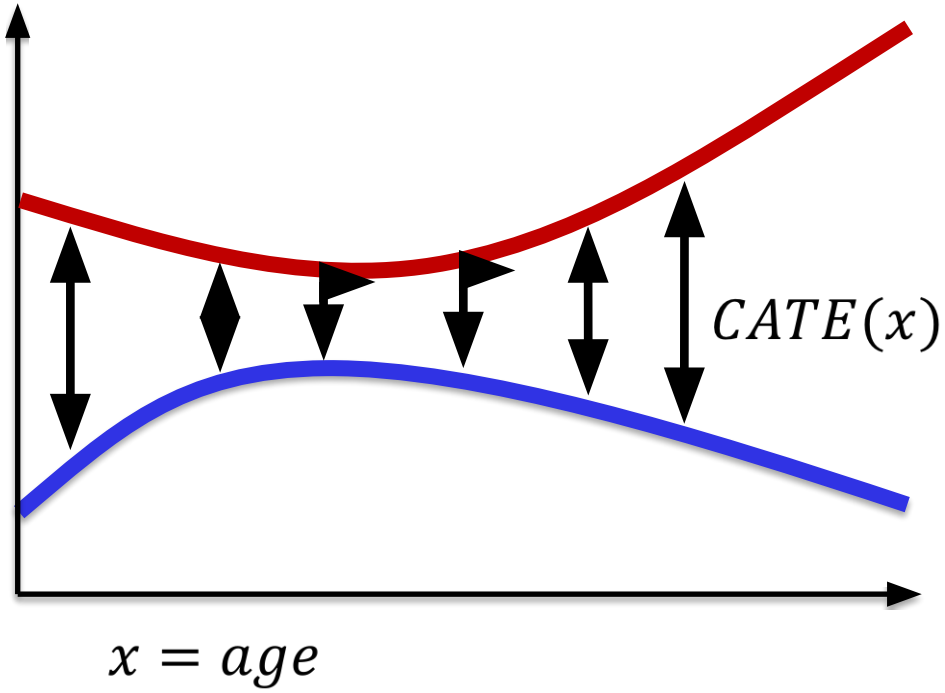


Blood Pressure and Age

$y = blood\_pres.$

$Y_1(x)$

$Y_0(x)$

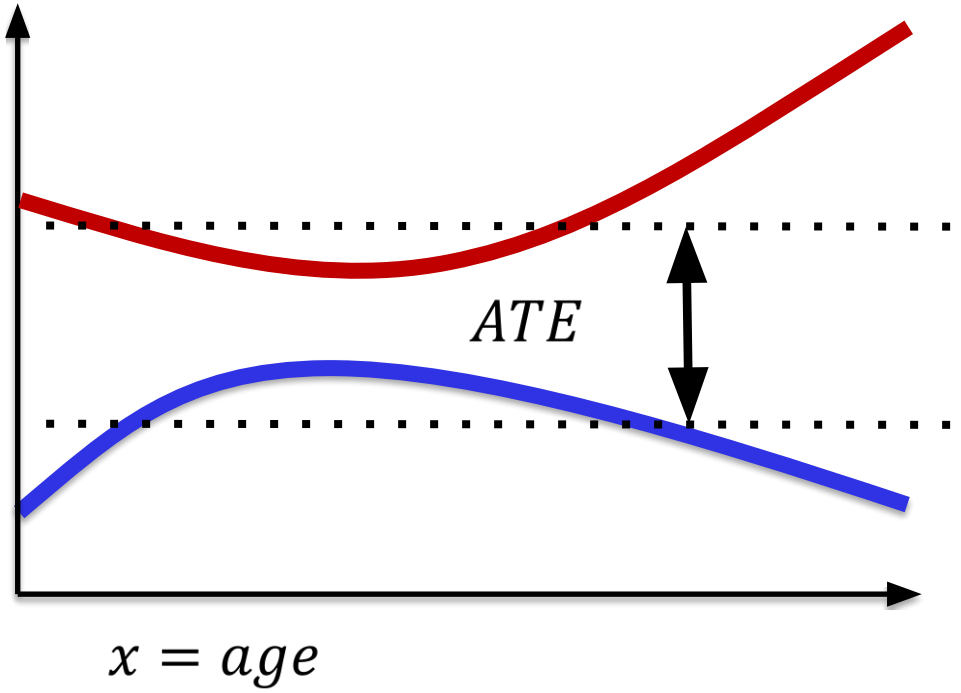


Blood Pressure and Age

$y = blood\_pres.$

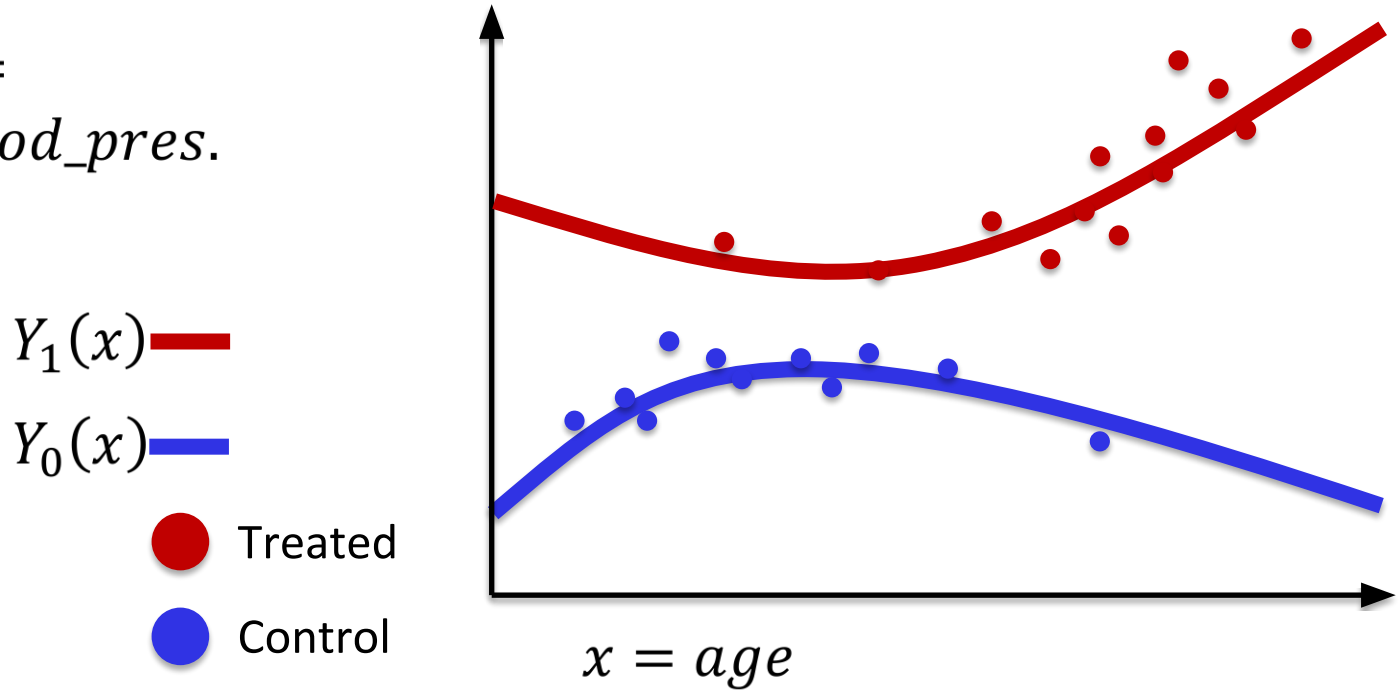
$Y_1(x)$

$Y_0(x)$



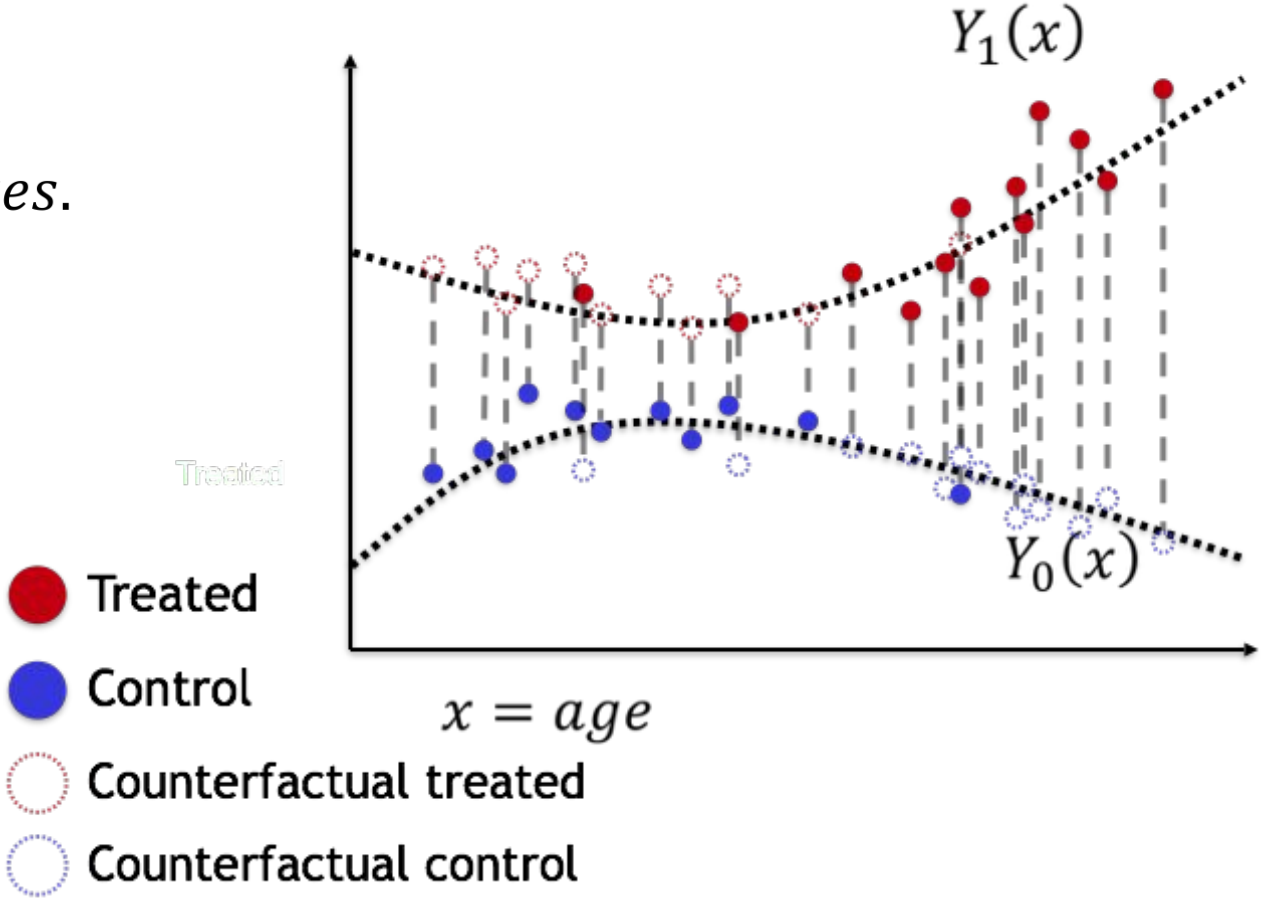
Blood Pressure and Age

$y = blood\_pres.$



Blood Pressure and Age

$y =$
$blood\_pres.$



“The fundamental problem of causal inference”

We only ever observe one of the two outcomes

Estimation Example

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

- True treatment effect:
  $\mathbb{E}[Y_1 - Y_0] = 2$

$\mathbb{E}[Y|t = 1] - \mathbb{E}[Y|t = 0] =$
$\frac{1}{4}(10 + 6 + 6 + 6) +$
$\frac{1}{4}(8 + 8 + 8 + 4) =$

$7 - 7 = 0$

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Estimation

- True treatment effect:
  $\mathbb{E}[Y_1 - Y_0] = 2$

$\mathbb{E}[Y|t = 1] = 7$
$\mathbb{E}[Y|t = 0] = 7$

$\mathbb{E}[Y|t = 0, Gender = M] = 8$
$\mathbb{E}[Y|t = 1, Gender = M] = 10$

$\mathbb{E}[Y|t = 0, Gender = F] = 4$
$\mathbb{E}[Y|t = 1, Gender = F] = 6$

Within each group we get the true treatment effect!

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment Assignment Mechanism

- G=0 if gender=F,
  G=1 if gender=M

  $Y_0 = 4+4*G$
  $Y_1 = 4+4*G+2$

- p(t=1|G=1) = 0.25
  p(t=1|G=0) = 0.75

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Random Treatment Assignments

They work because it allows to get expectations from observations!

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1|T = 1] =$
- $\mathbb{E}[Y_{obs}|T = 1]$

  Can be estimated from data

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0|T = 0] =$
- $\mathbb{E}[Y_{obs}|T = 0]$

Can be estimated from data

Completely Random Treatment Assignments

They work because it allows to get expectations from observations!

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_1] =$
- $\mathbb{E}[Y_1 | T = 1] =$
- $\mathbb{E}[Y_{obs} | T = 1]$

Can be estimated from data

- Treatment is random: $(Y_0, Y_1) \perp\!\!\!\perp T$
- $\mathbb{E}[Y_0] =$
- $\mathbb{E}[Y_0 | T = 0] =$
- $\mathbb{E}[Y_{obs} | T = 0]$

Can be estimated from data

$$ATE = \mathbb{E}[Y_1 - Y_0] = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y_{obs} | T = 1] - \mathbb{E}[Y_{obs} | T = 0]$$

Note the difference because unobservable quantities (potential outcomes) and observable quantities

Treatment Assignment Is Not Random!

Gender	Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment Assignment Is Not Random!

$$P(Y_0 = 8|T = 0) = 0.75$$
$$P(Y_0 = 8|T = 1) = 0.25$$
$$P(Y_1 = 10|T = 0) = 0.75$$
$$P(Y_1 = 10|T = 1) = 0.25$$

(Y_0, Y_1) ***are not*** independent of T

Gender	T: Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

Treatment Assignment Is Not Random!

$$P(Y_0 = 4|T = 0, G = F) = 1$$
$$P(Y_0 = 4|T = 1, G = F) = 1$$
$$P(Y_1 = 6|T = 0, G = F) = 1$$
$$P(Y_1 = 6|T = 1, G = F) = 1$$

(Y_0, Y_1) ***are independent*** of T ***conditioned*** on G=M, and conditioned on G=F

$$(Y_0, Y_1) \perp\!\!\!\perp T|\mathrm{G}$$

Gender	T: Treatment	Y_0: Sugar levels *had they received* treatment 0	Y_1: Sugar levels *had they received* treatment 1	Y: Observed sugar levels
M	0	8	10	8
M	0	8	10	8
M	0	8	10	8
M	1	8	10	10
F	0	4	6	4
F	1	4	6	6
F	1	4	6	6
F	1	4	6	6

No Unmeasured Confounding!

If We Cannot Randomize Treatment?

- We can still succeed if the treatment assignment process is *conditionally randomized*, conditioned on an observed quantity.

- This is actually just a way of saying we have *no unmeasured confounding.*

“The Assumptions”

Sufficient conditions for us to identify the causal effect in an observational study?

- No unmeasured confounders
- Common support

Ignorability - No Unmeasured Confounding

$$(Y_0, Y_1) \perp\!\!\!\perp T \mid x$$

The potential outcomes are independent of treatment assignment, conditioned on observed covariates x

Failure: In the example above, gender was associated with the potential outcomes **and** treatment assignment

Unverifiable from data!

Common Support Assumption

- Y_0, Y_1: potential outcomes for control and treated

  x: unit covariates (features)

  T: treatment assignment

  We assume:

$$p(T = t|X = x) > 0 \; \forall t, x$$

Example: Running an observational study

- Check your assumptions and design!
- Is there reason to believe no unmeasured confounding holds? Use domain knowledge
- More generally, do you believe ignorability holds?
- If not - change the design:
  - Add more variables
  - Measure treatment differently
  - Measure outcome differently

Example: Running an observational study

Comparing effectiveness of two anti-hypertensive medications:

- Treatment: first administration of medication
- Outcome: blood pressure 3 months after first treatment

  But is outcome only measured for some of the patients?

- Did we measure the important known causes of hypertension? Literature survey may reveal that high alcohol use is a known cause of hypertension

- Doctors know this, and might use this information in deciding on treatment

- If we don't measure alcohol use, it becomes hidden confounder which might bias our conclusions

Check for Overlap

- Check for overlap between treated and control on important univariate and bivariate variables, e.g. age, gender, weight in a medical study

- If no overlap, redefine study population, e.g. only people ages 40-60

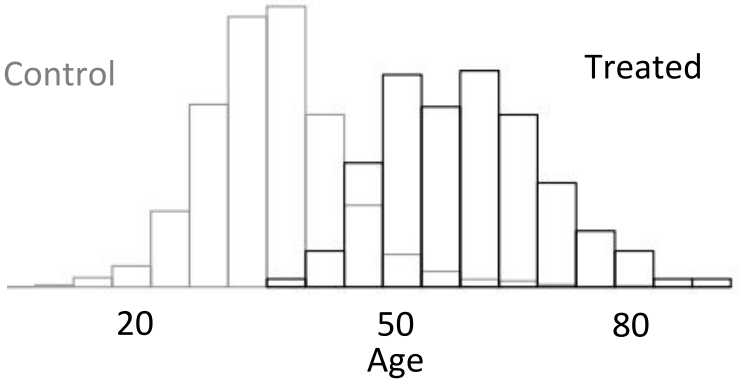
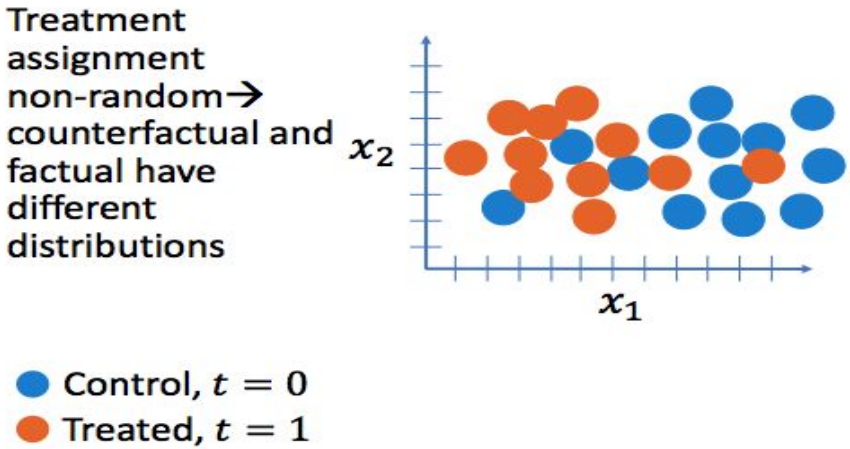


Figure:
Hill & Gelman

What Else Can We Use? Propensity Score!

- Extremely widely used tool
- Basic idea: turn observational study into a pseudo-randomized trial by correcting for non-random sampling



Propensity Score

- $(Y_0, Y_1) \perp\!\!\!\perp T \mid x$
- What functions of $f(x)$ will still allow $(Y_0, Y_1) \perp\!\!\!\perp T \mid f(x)$?
- Theorem:
  Let $e(x) = p(T = 1|x)$, also called the ***propensity score.***
  If ignorability holds for x, then $e(x)$ is the coarsest function of x for which ignorability still holds
- If we have ignorability, in theory the propensity score gives us everything we need
- We can run covariate adjustment on the propensity score!

$$\mathbb{E}[Y|e(x), T = 1] - \mathbb{E}[Y|e(x), T = 0]$$

- Other methods using propensity which we will see soon:
  - Inverse propensity score weighting
  - Propensity score matching
  - Stratification on the propensity score

Propensity Score

- $e(x) = p(T = 1|x)$, the treatment assignment mechanism
- In most cases must be estimated from data
- Can use any machine learning method:
  logistic regression, random forests, neural nets
- Unlike most ML applications, we need to get the **probability** itself accurately
- Subtle point: if we include x which are only predictive of treatment assignment but not outcome
- Hard (but not impossible) to validate models

Propensity Score - Algorithm for ATE Estimation

- How to calculate ATE with propensity score for sample $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$

1. Use any ML method to estimate $\hat{p}(T = t|x)$
2. $$\hat{ATE} = \frac{1}{n} \sum_{i \text{ s.t. } t_i=1} \frac{y_i}{\hat{p}(t_i = 1|x_i)} - \frac{1}{n} \sum_{i \text{ s.t. } t_i=0} \frac{y_i}{\hat{p}(t_i = 0|x_i)}$$

Not Covered: Propensity Score Matching

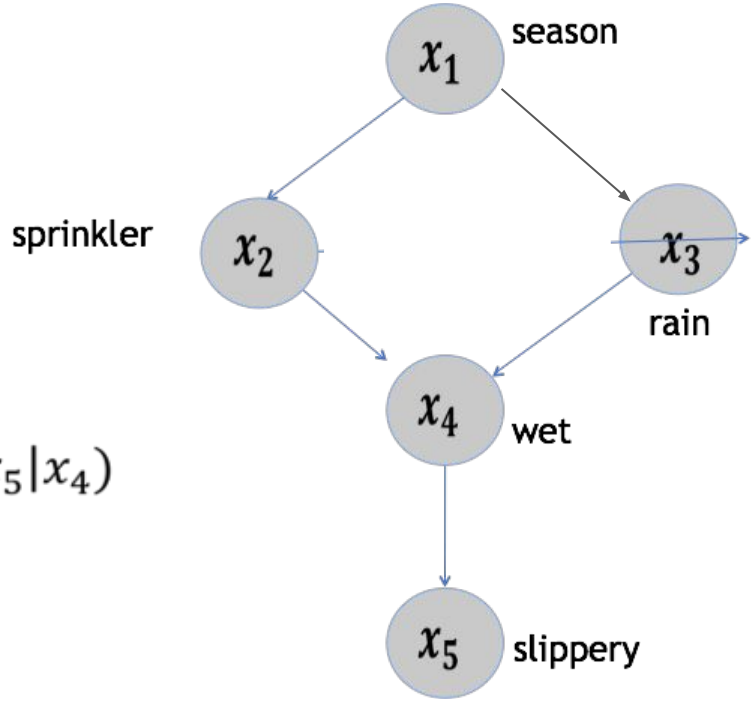
Outline

Slides today are courtesy of Shalmali Joshi and Uri Shalit!

1. What is confounding?
2. Why causal reasoning?
3. Potential Outcomes and Propensity Scoring
4. **Pearlean Causal Graphs Framework**
5. Project ideas

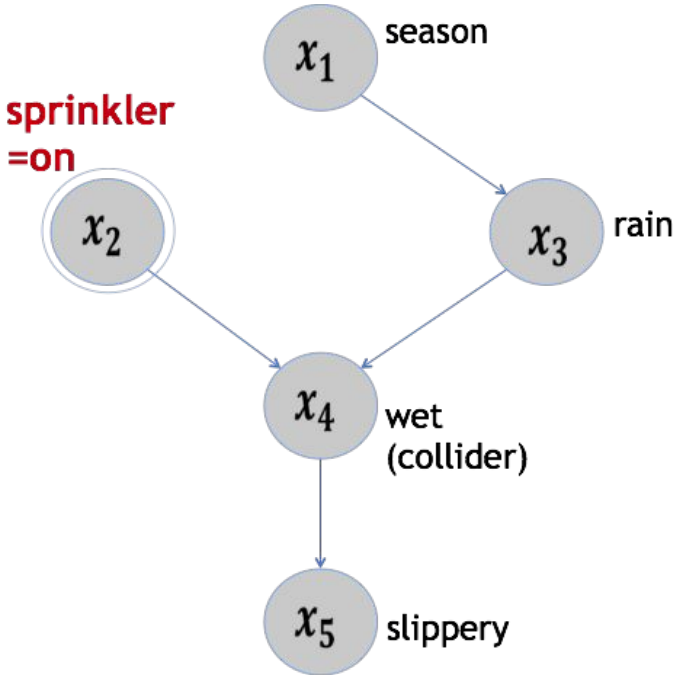
Pearlean Causal Framework

$$p(x_1, x_2, x_3, x_4, x_5) =$$
$$p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_3, x_2)p(x_5|x_4)$$



Intervention

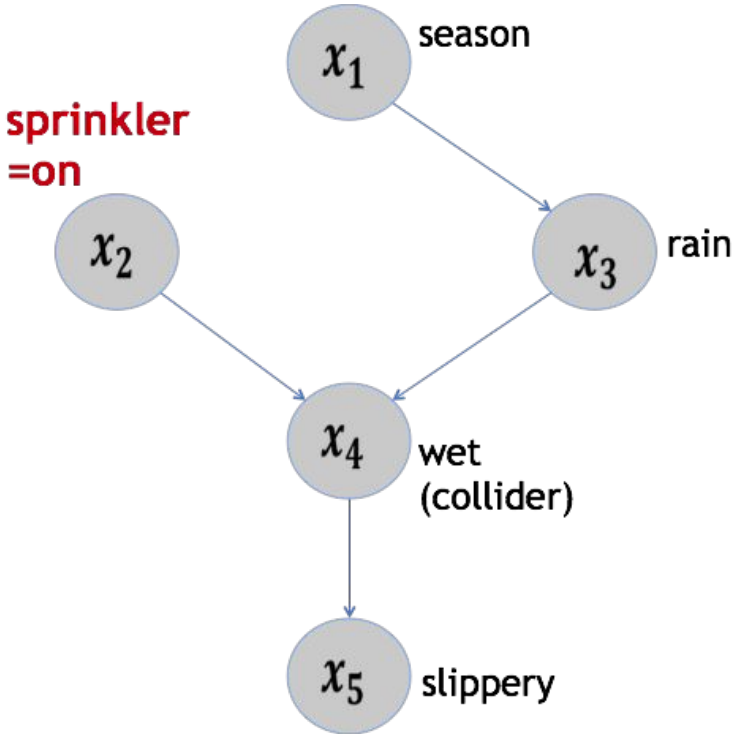
- Turn the sprinkler on, please
- We removed the association between season and sprinkler
- We are now in a new world, where the sprinkler is set to on
- This is the do-operator



Intervention (do-Calculus)

- $p_{do(x_2=on)}(x_1, x_3, x_4, x_5) = p(x_1)p(x_3|x_1)p(x_4|x_3, x_2 = on)p(x_5|x_4)$

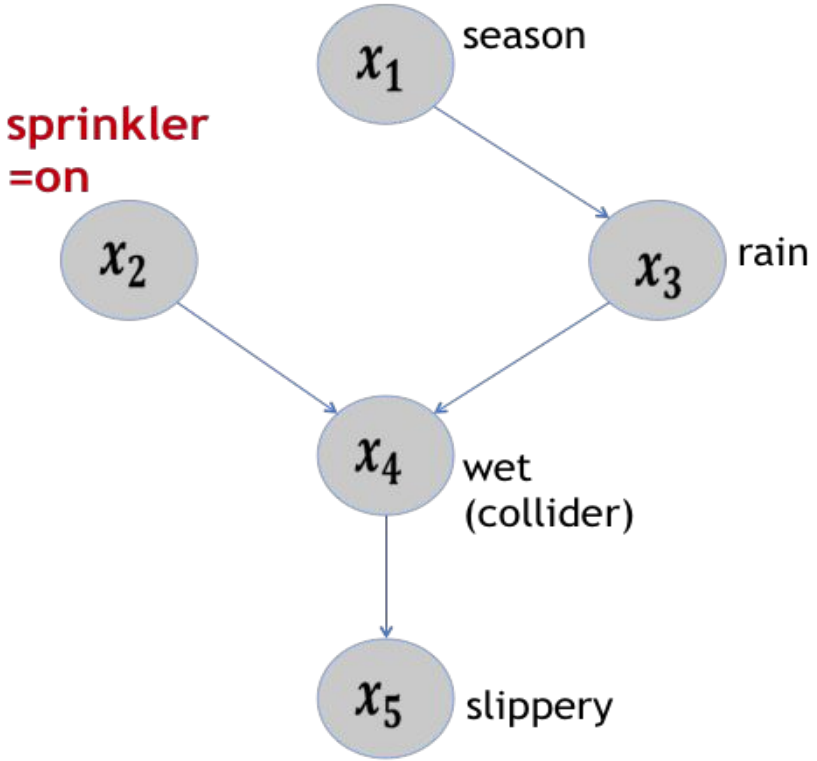
- $p(x_1, x_3, x_4, x_5|x_2 = on) = p(x_1|x_2 = on)p(x_3|x_1, x_2 = on) \cdot p(x_4|x_3, x_2 = on)p(x_5|x_4, x_2 = on)$



do-operator versus conditioning

- $p(x_1, x_3, x_4, x_5 | do(x_2) = on)$
  distribution under an ***action***

- $p(x_1, x_3, x_4, x_5 | x_2 = on)$
  distribution given ***evidence***



The Assumptions: Causal Identifiability

- Can we infer $p(y|do(v))$ from some observed $p(y, v, x)$?
- If there are $p_1(y|do(v)) \neq p_2(y|do(v))$ that are both consistent with $p(y, v, x)$ then the answer is no
- How can we tell if $p(y|do(v))$ is uniquely determined by $p(y, v, x)$?
- Causal graphs give us many different sufficient conditions
- Without knowing the causal graph, the same observable distribution can result from two very different causal processes
- Very different conclusions about which treatment we should use
- Causal graphs can give us sufficient conditions for when causal queries $p(y|do(v))$ are identifiable from an observed distribution
- Causal graphs encode extra knowledge!

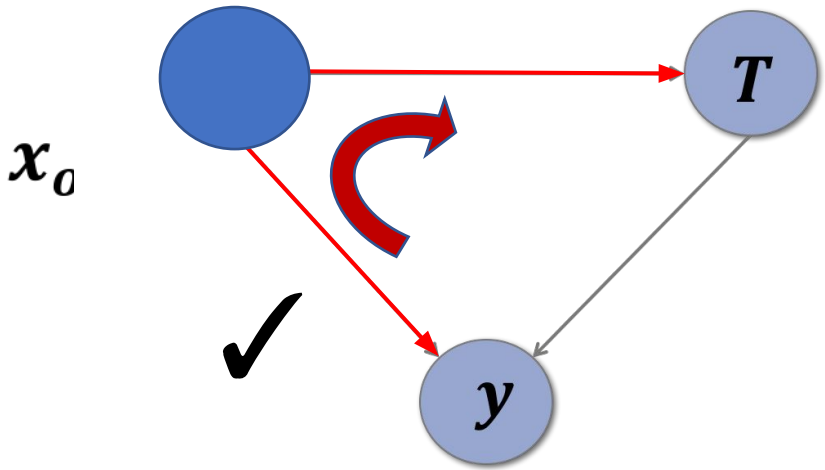
Backdoor Criteria

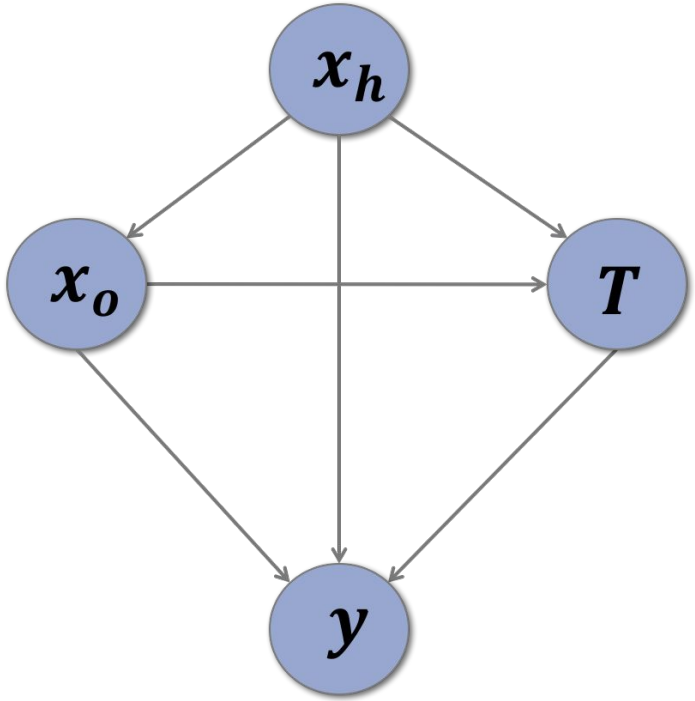
- Back-door criterion (Pearl, 1993, 2009):
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T
- Tells us what can we measure that
  will ensure causal identifiability

- A set of variables Z satisfies the back-door criterion relative to the ordered pair (T,Y) if:
  1. No node in Z is a descendant of T; and
  2. Z blocks (in the d-separation sense) every path between T and Y that contains an arrow into T

The Assumptions: Causal Identifiability

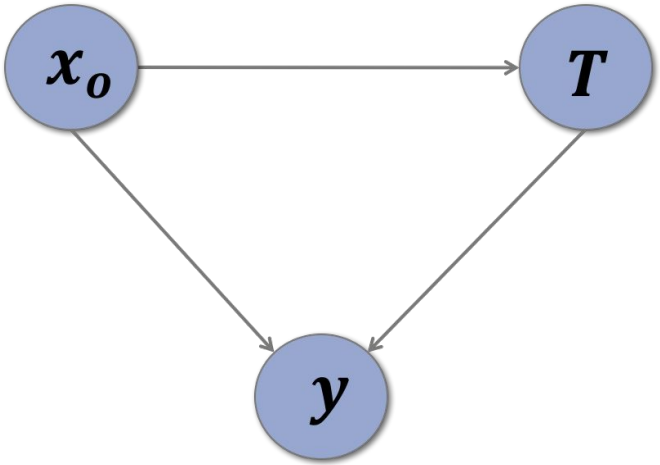
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

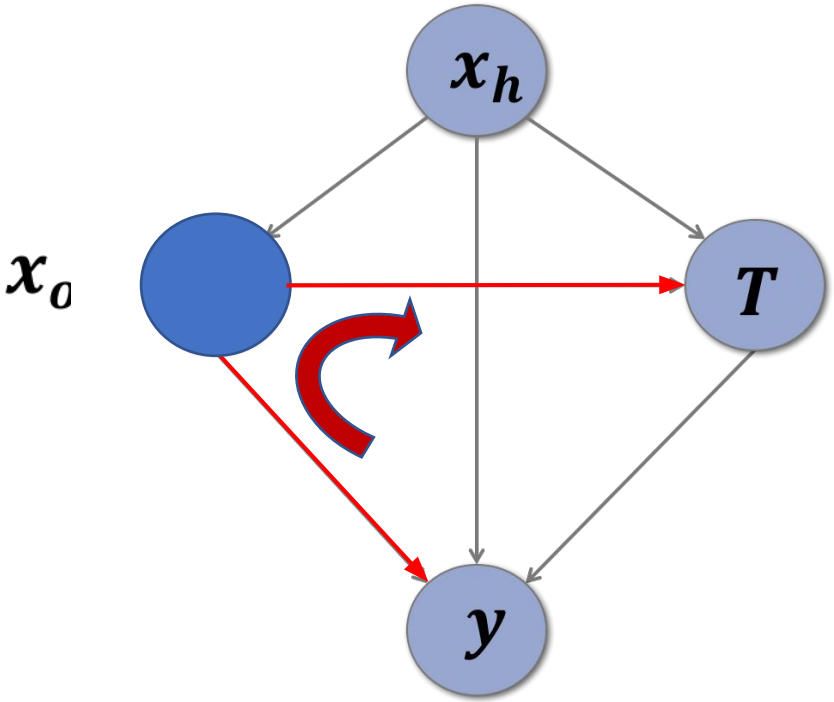




The Assumptions: Causal Identifiability

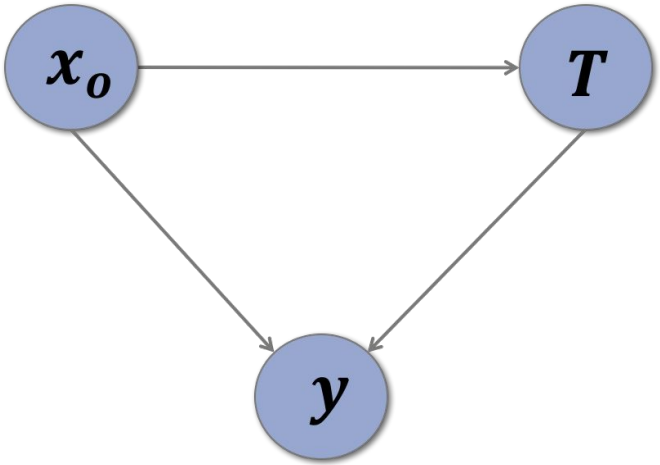
- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

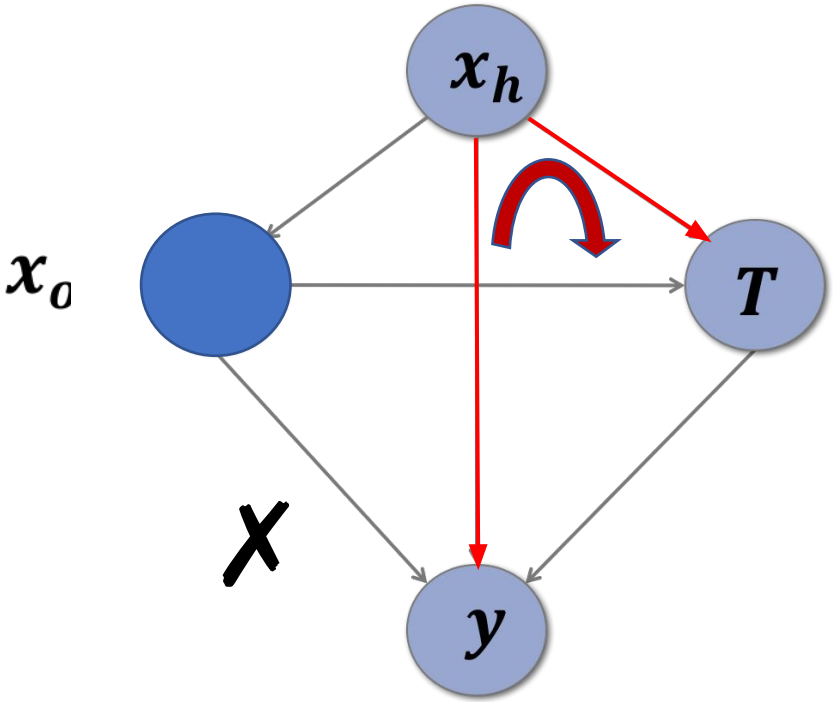




The Assumptions: Causal Identifiability

- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

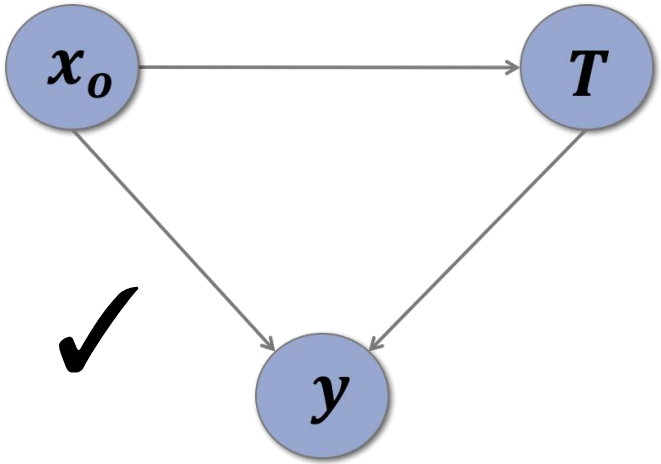




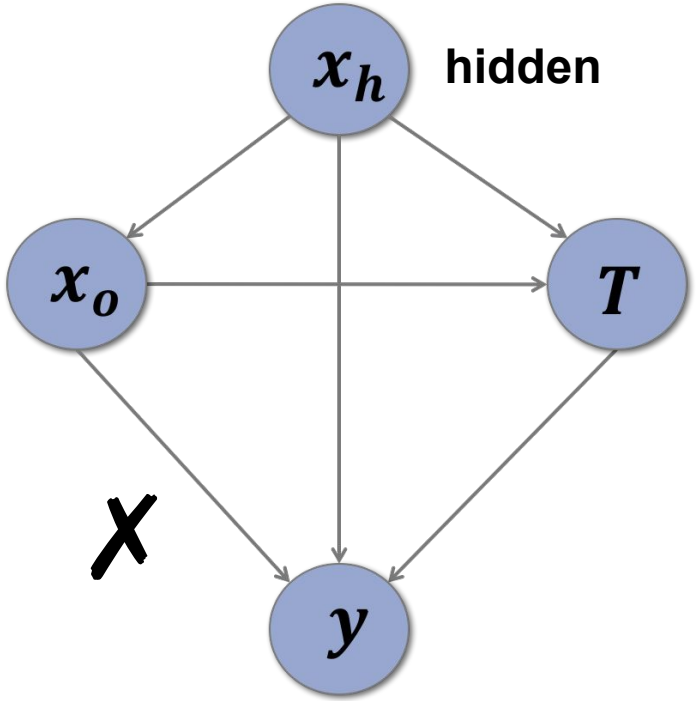
The Assumptions: Causal Identifiability

- Back-door criterion:
  The observed variables d-separate all paths between y and T that end with an arrow pointing to T

x_o

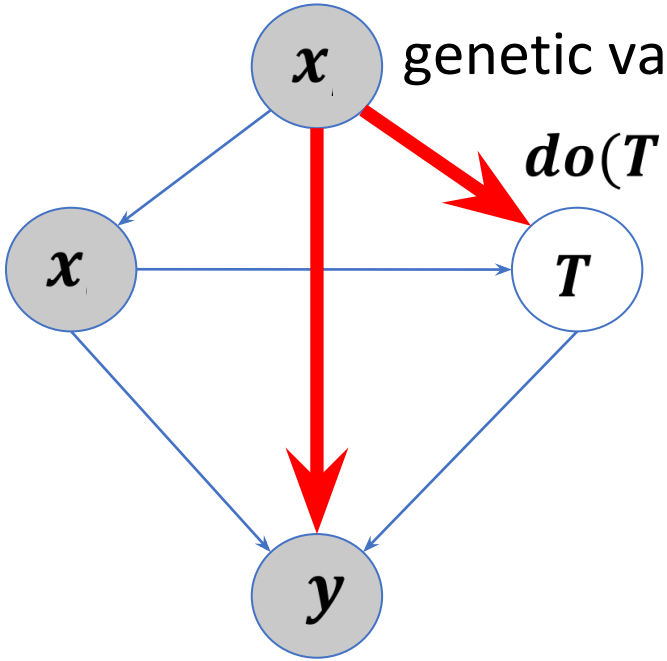


✓



✗

Unidentifiable Causal Effect



genetic variables

$do(T = t)$

age, gender, weight, diet, heart rate at rest,…

anti-hypertensive medication

post-treatment blood pressure

Simpson's Paradox

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

- Why is this a paradox? "When my parts are summed, am I less than some of my parts?"
- Should doctors focus on gender? Can you spot the lurking variable?

Main Takeaways

- Supervised learning has limitations
- RCTs are expensive AND limited; think causally especially for clinical data
- Pearl’s and Rubin’s frameworks provide foundational formalism for causal effect estimation
- Not all effects are identifiable
- Most research questions cater to how to relax all the assumptions we made along the way!

Course Reminders!

- Submit the weekly reflection questions to MarkUs!
- Start the homework!

  Q/A session on the problem sets
  Wednesday, Jan 23 at 4-6pm in GB 405

  Monday, Feb 4 at 4-6pm in SS 1071
- Sign up for a paper presentation slot!
- Think about your projects!