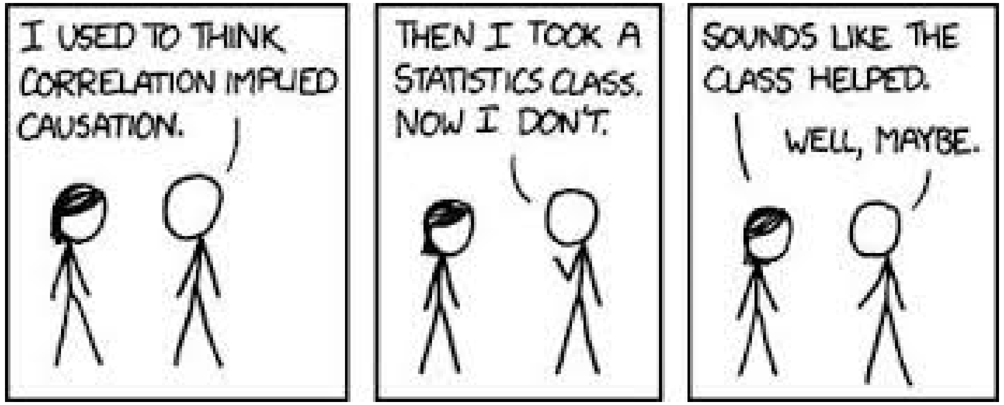
CS 189/289

One-class intro to “causality”

1. Some intuition
2. Some formalism



This lecture is based in part on notes from Prof. Moritz Hardt.
For more details, see Chapters 9 & 10 here: https://mlstory.org/

CS 189/289

One-class intro to “causality”

1. Some intuition
2. Some formalism

ML prediction: causation or correlation?

- So far: take observed data, $D = \{x_i, y_i\}$; propose a model class, $\hat{y}_i = f_\theta(x_i) = p_\theta(y|x)$
- MLE to obtain $\hat{\theta}$.
- Suppose get 99% accuracy with cross-validation.
- Is $p_\theta(y|x)$capturing the underlying *causes* of y?
- Does it matter?

Actual *vs.* predicted sale price of house

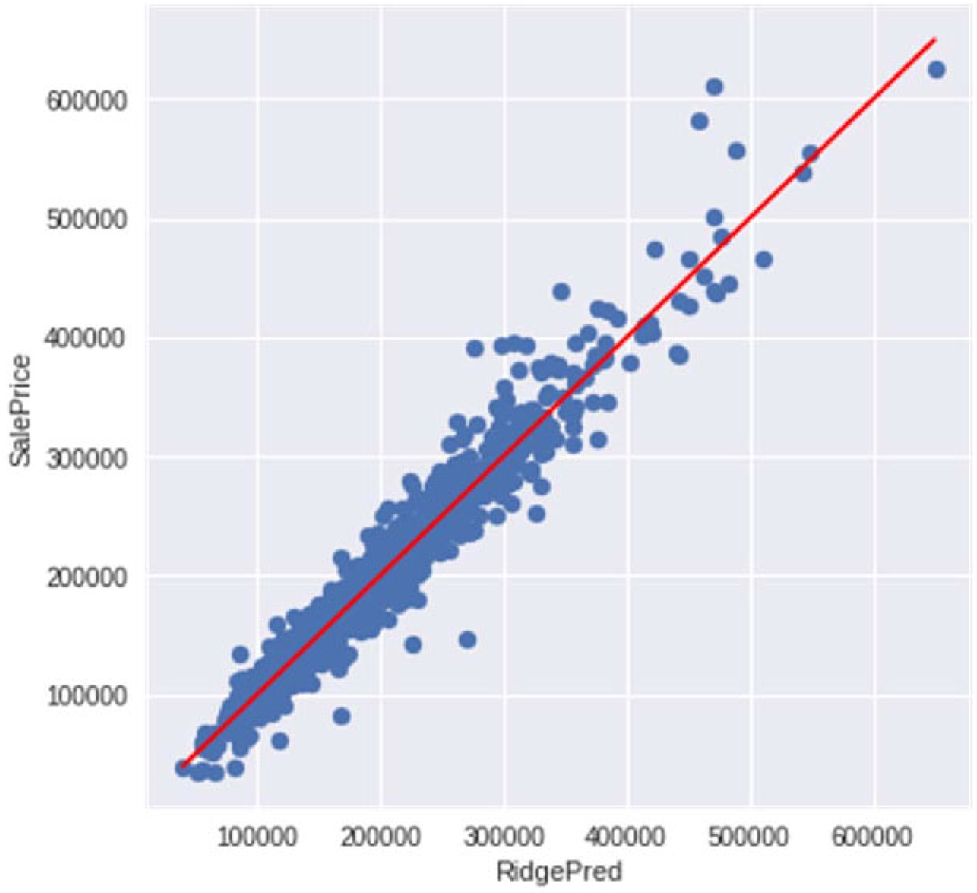


Fig. 4 Ridge Prediction for Training Data.

ML prediction: causation or correlation?

Breakingviews

Zillow’s failed house flipping

Reuters

WSJ NOV. 2021 : *“The company expects to record losses of more than $500 million from home-flipping by the end of this year and is laying off a quarter of its staff.”*

Actual *vs.* predicted sale price of house

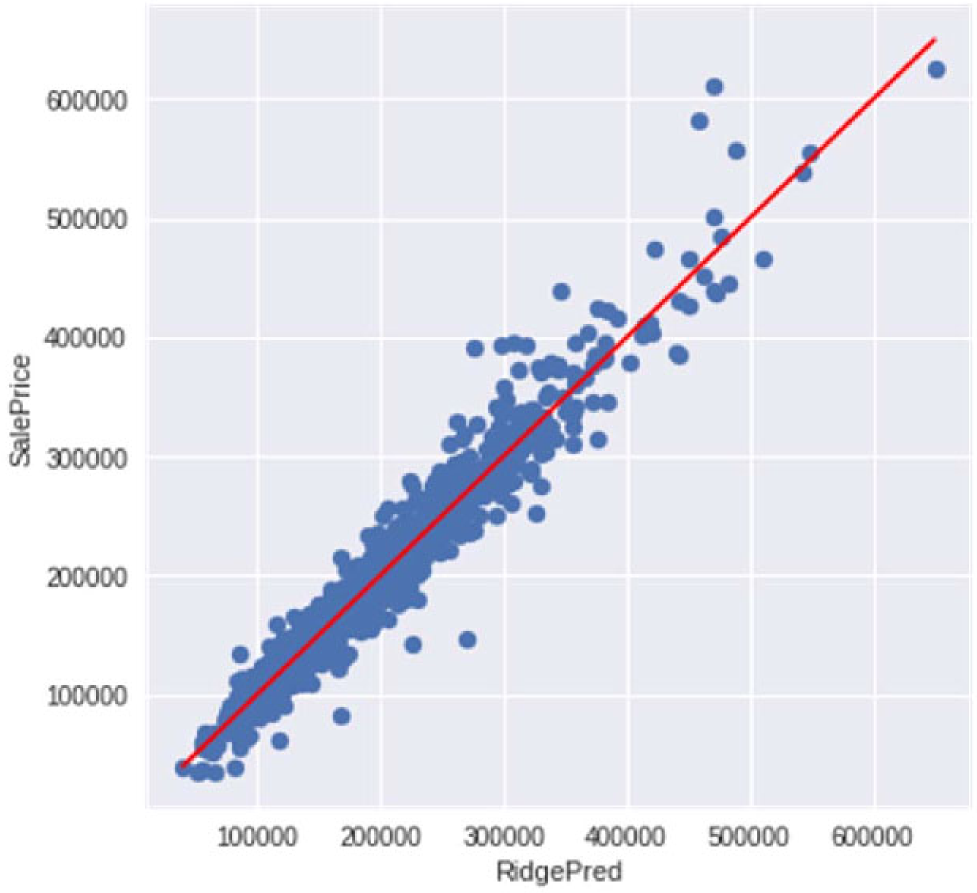
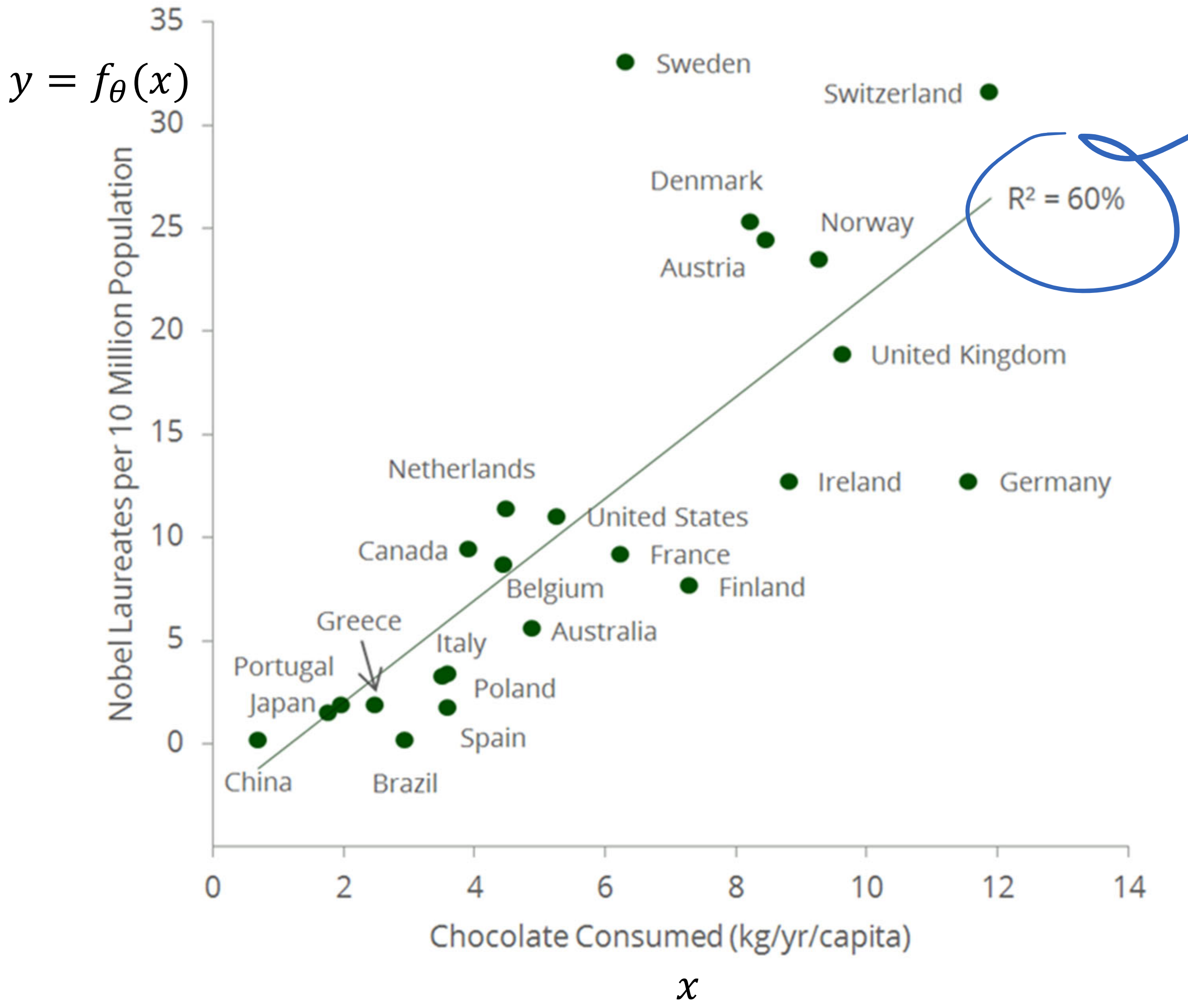
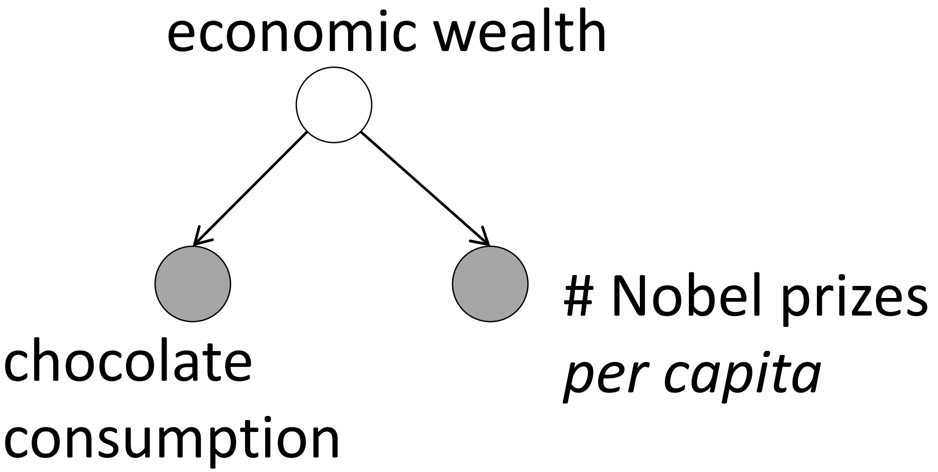


Fig. 4 Ridge Prediction for Training Data.

ML prediction: causation or correlation?



- This is a consequence of economic wealth.
- Richer countries spend more on education and luxury goods, like chocolate.



ML prediction: causation or correlation?

The Washington Post

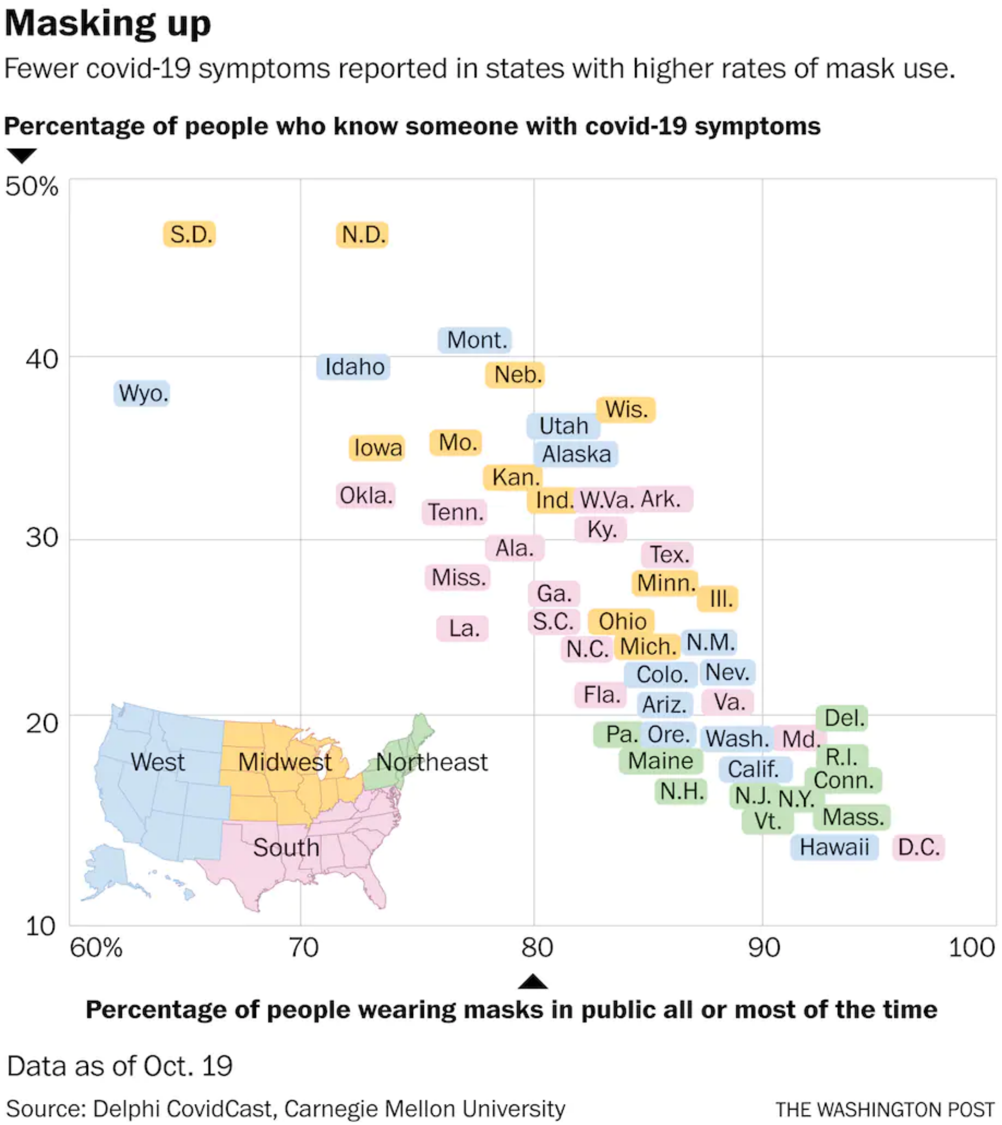
By Christopher Ingraham
Reporter

October 23, 2020

Business • Analysis

A ~~powerful~~ incorrect argument for wearing a mask, in visual form

Real-time pandemic data paints a vivid picture of the relationship between mask-wearing and the prevalence of covid-19 symptoms



A classic conundrum: kidney stone treatment

- Effectiveness of treatments A vs B for kidney stones from hospital data.
- Goal: is treatment A or B better?

Treatment A	Treatment B
273/350 (78%)	289/350 (83%)

higher success rate

A classic conundrum: kidney stone treatment

- Effectiveness of treatments A vs B for kidney stones from hospital data.
- Goal: is treatment A or B better?

Size of stones	Treatment A	Treatment B	#	% assigned B
All sizes	273/350 (78%)	289/350 (83%)		
Large stones	192/263 (73%)	55/80 (69%)	263+80=343	80/343=23%
Small stones	81/87 (93%)	234/270 (87%)	87+270=357	270/357=76%

[*Charig et al. BMJ 1986*]

Huh? What is going on? Which treatment would you want?

Possible explanation: doctors assign B more often to small stones, which are easier to treat.

A classic conundrum: kidney stone treatment

- Effectiveness of treatments A vs B for kidney stones from hospital data.
- Goal: is treatment A or B better?

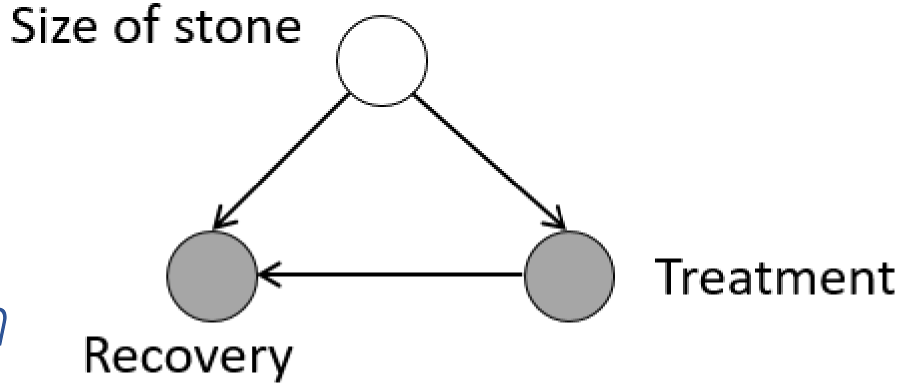
Size of stones	Treatment A	Treatment B
All sizes	273/350 (78%)	289/350 (83%)
Large stones	192/263 (73%)	55/80 (69%)
Small stones	81/87 (93%)	234/270 (87%)

[Charig et al. BMJ 1986]

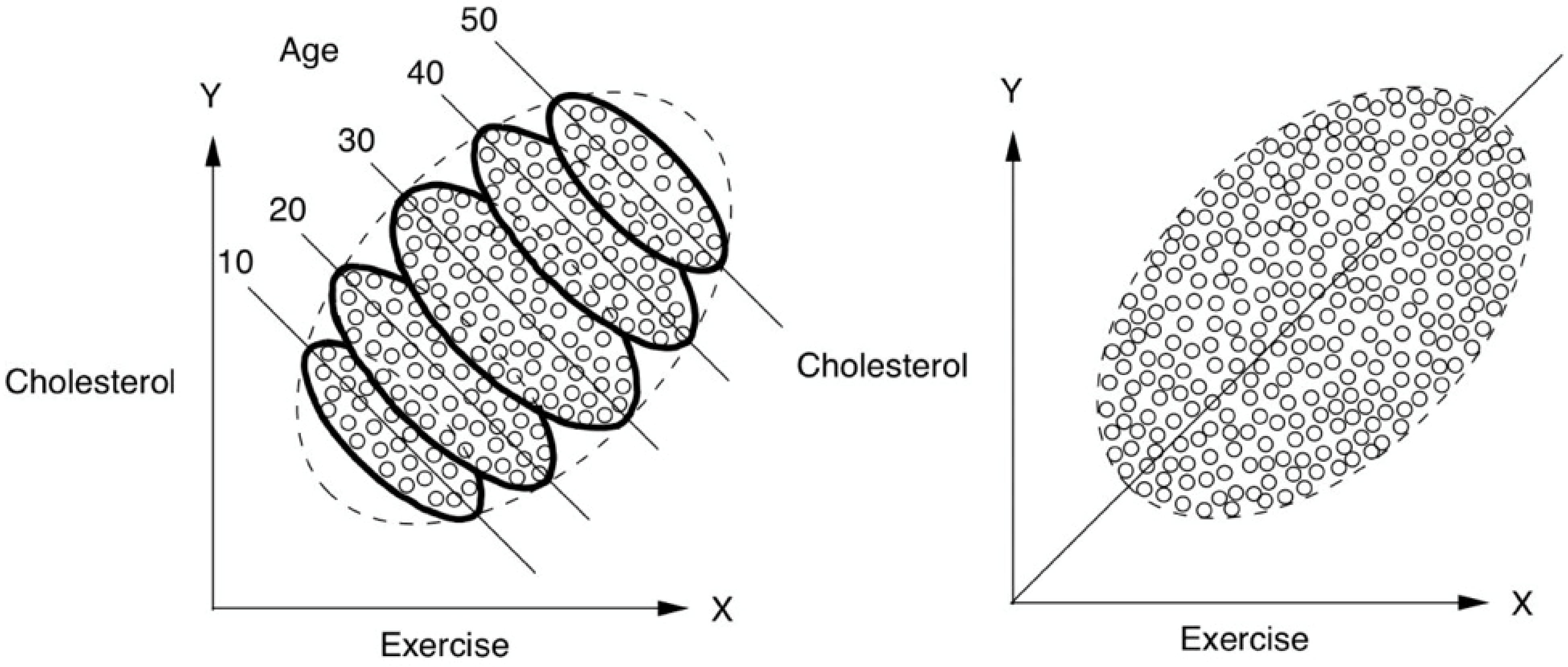
Huh? What is going on? Which treatment would you want?

Possible explanation: doctors assign B more often to small stones, which are easier to treat.

- This is an example of *Simpson's paradox*.
- With a more careful understanding, it is not really paradoxical.
- The stone size is a *confounding* variable:



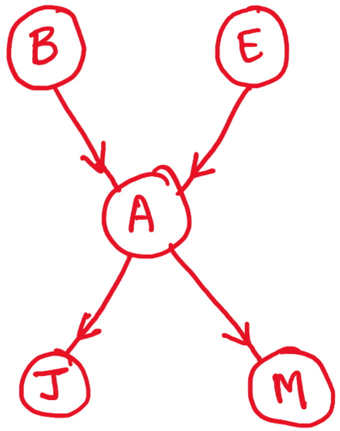
One visualization of Simpson's Paradox



[The book of why: the new science of cause and effect, Judea Pearl and Dana MacKenzie.]

Are *probabilistic graphical models* causal models?

- Previously you learned about *probabilistic graphical models*, like HMMs.
- In general, these models do not reason about causes, nor speak to causality.
- With additional assumptions, we can leverage the machinery of graphical models to reason about causality: *Structural Equation Models* (SEM) (soon).



$P(B, E, A, J, M) =$
$P(B)\,P(E)\,P(A \mid B, E)\,P(J \mid A)\,P(M \mid A)$

There are 2^5 entries in the joint probability distribution

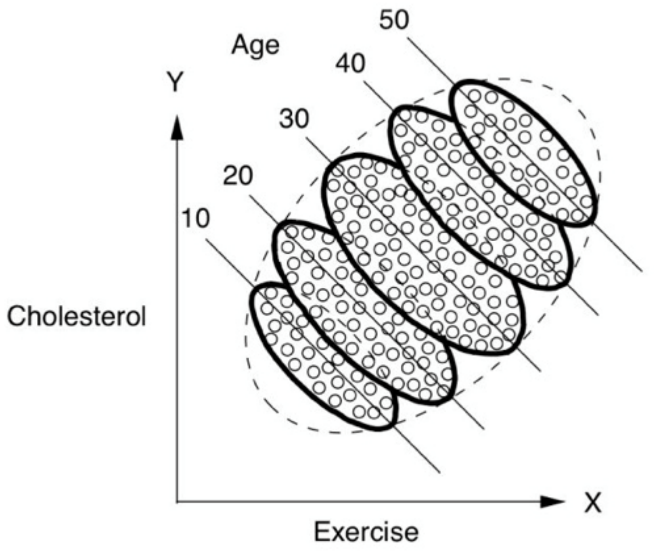
This "factorized" representation makes it much more concise.
10 numbers instead of 31

More on Simpson's Paradox

Formally, the paradox can be stated as follows:

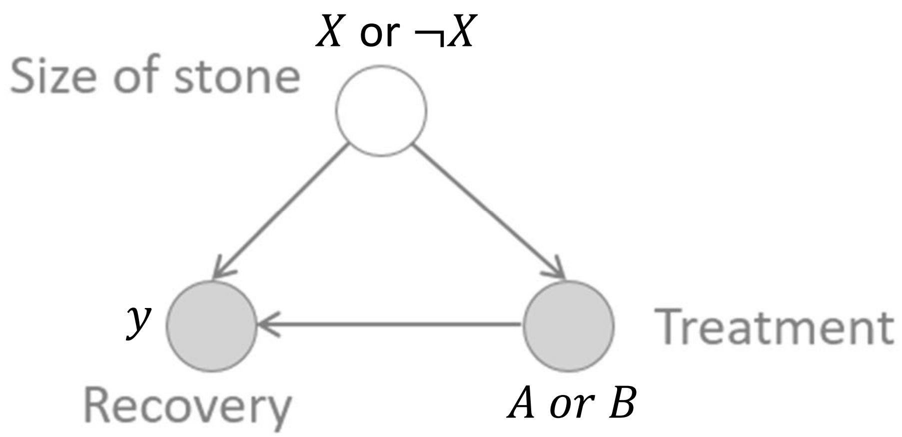
1. $p(y|A) \quad < p(y|B)$ ("All sizes")
2. $p(y|A,X) \quad > p(y|B,X)$ ("Large stones")
3. $p(y|A,\neg X) > p(y|B,\neg X)$ ("Small stones:)

probability of recovery



Size of stones	Treatment A	Treatment B
All sizes	273/350 (78%)	289/350 (83%)
Large stones	192/263 (73%)	55/80 (69%)
Small stones	81/87 (93%)	234/270 (87%)

[Charig et al. BMJ 1986]



Revisiting Simpson's Paradox

Formally, the paradox can be stated as follows:

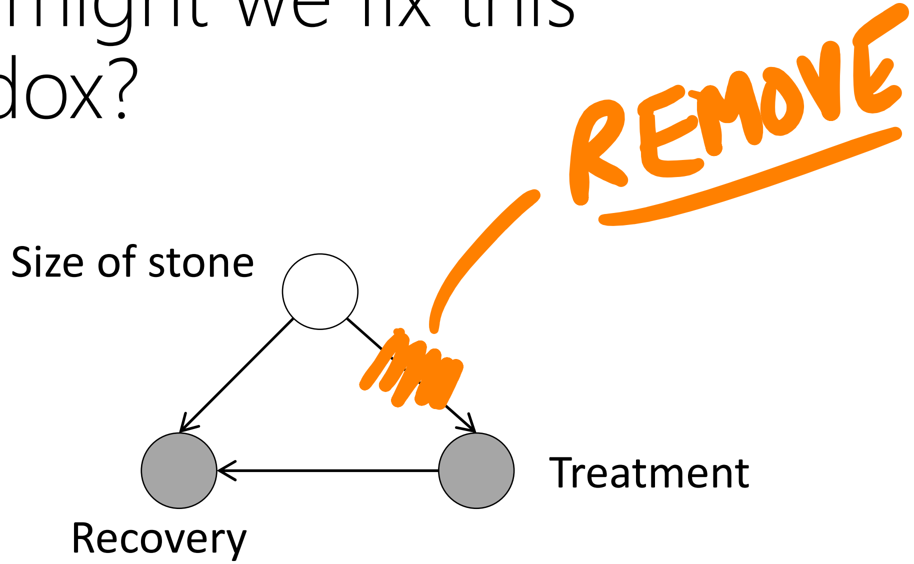
1. $p(y|A) < p(y|B)$ ("All sizes")
2. $p(y|A, X) > p(y|B, X)$ ("Large stones")
3. $p(y|A, \neg X) > p(y|B, \neg X)$ ("Small stones:)

- Mathematically, no contradiction, so why the seeming paradox?
- We tend to interpret conditional events as *actions*, but they are not.
- Conditional events are *observations*.
- We'll learn more.

Revisiting Simpson's Paradox

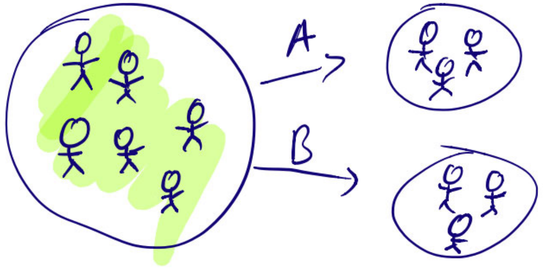
- We observe doctors in a hospital.
- *i.e.*, we see who gets treatment A or B, according to the doctor's internal decisions system ("natural inclination").
- There is no intervention (no action), just passive observation.
- If we could redesign the experiment, how might we fix this problem so that we avoid Simpson's paradox?

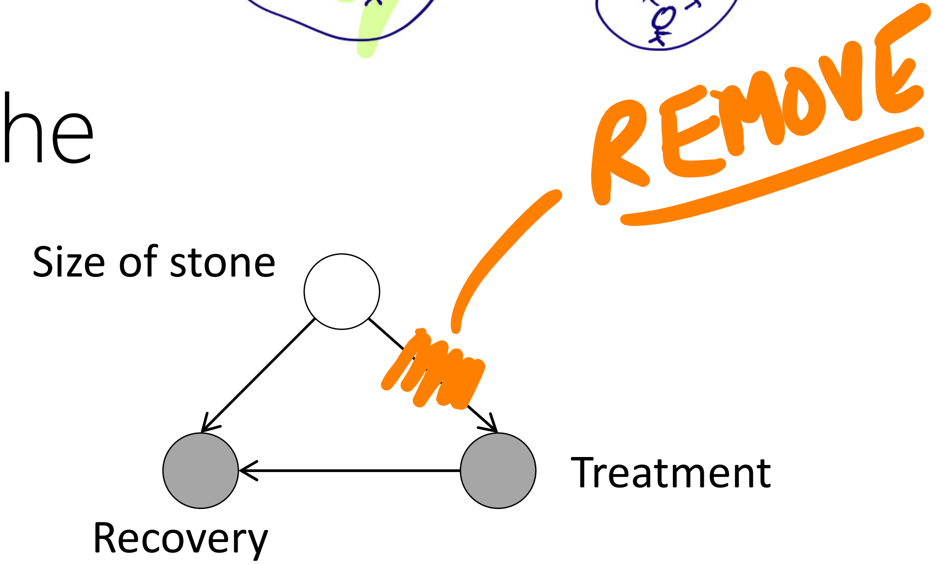
Size of stones	Treatment A	Treatment B
All sizes	273/350 (78%)	289/350 (83%)
Large stones	192/263 (73%)	55/80 (69%)
Small stones	81/87 (93%)	234/270 (87%)



Randomized Controlled Trial (RCT)

- Are the "gold standard" way to conduct such experiments.
- i.e., *disallow the use of the doctor's* internal decision system in assigning the choice of treatment.
- Replace the doctor's decision with one created at random—we *act* on the system.
- Now the doctor cannot more frequently assign Treatment B to the smaller stones.

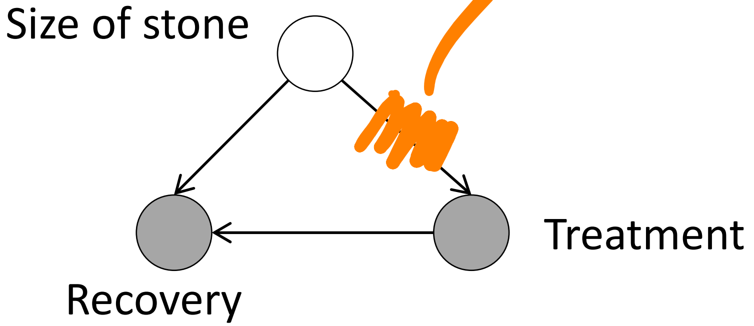




Randomized Controlled Trial (RCT)

- This is the difference between an *observational* and a *randomized* experiment.
- The randomization process is called an *intervention* (or action) in the field of causality.
- It is easier to extract causality using interventional data than using observational data.

REMOVE

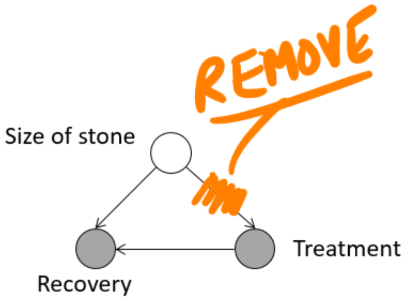


CS 189/289

One-class intro to “causality”

1. Some intuition
2. Some formalism

How do we formalize *actions*?



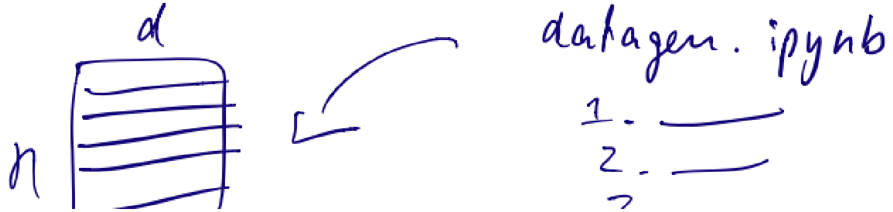
- We saw how *intervening* on *upstream causes* of the treatment variable could eliminate the confounding variable.
- Actions are not conditional events, so we need a new notation/concept beyond $p(y|A, X)$.
- The "do" action notation looks like conditional probabilities, but isn't, $p(y = 1|do(A = 1))$.
- We'll discuss the relationship between these.

How do we formalize *actions*?

- We are going to work our way toward the formalism of *Structural Equation Models* (SEMs).
- SEMs are equivalent to defining a *causal data-generating process*.
- i.e., think of SEMs as writing code that would generate the data, step-by-step, through each causal mechanism.

Data vs. source code to generate it?

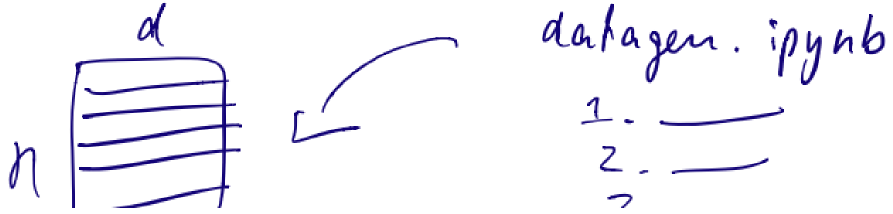
- Suppose someone asks you to help them understand some data they have.
- They ask if you would prefer to have the source code that generated it, or just the data itself. Which would you prefer?



Data vs. source code to generate it?

- Suppose someone asks you to help them understand some data they have.
- They ask if you would prefer to have the source code that generated it, or just the data itself. Which would you prefer?
- The code contains more information (we can generate data from the program, but not the other way around).

E.g. "does A cause y ?"



- Also, we can change the code and generate different data, seeing which variables have effects on which other variables.

Programming intuition example (SEM)

Suppose you have a program to generate a distribution, step-by-step:

1. Sample Bernoulli random vars
   $U_1 \sim B(1/2)$, $U_2 \sim B(1/3)$, $U_3 \sim B(1/3)$
2. $X := U_1$ (exercise)
3. $W :=$ if $X=1$ then 0 else U_2 (overweight)
4. $H :=$ if $X=1$ then 0 else U_3 (heart disease)

- This induces a joint distribution over the binary RVs, X, W, H.
- We can compute various probabilities of potential interest:

$$p(H=1) = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

$$p(H=1|W=1) = \frac{1}{3}$$

- Thus $p(H=1|W=1) > p(H=1)$. Does this mean W causes H?
- If it did, then intervening/acting on W would change H. Lets try it.

Programming intuition example (SEM)

- In this new program, the probability of $H = 1$ is still equal to $\frac{1}{6}$.
- We write this as $p(H = 1|do(W = 1)) = \frac{1}{6}$, and call it a *do-intervention*.
- Generally $p(H = 1|do(W = 1)) \neq p(H = 1|W = 1)$

Suppose you have pr… distribution, step-by-s…

1. Sample Bernoulli ra…
   $U_1 \sim B(1/2)$, U_2 …
2. $X := U_1$ (exercise)
3. $W := 1$ (overweight)
4. $H :=$ if $X = 1$ then 0 else U_3 (heart disease)

"DOING" vs. "OBSERVING"

…interest.

$p(H = 1) = \frac{1}{2}\frac{1}{3} = \frac{1}{6}$

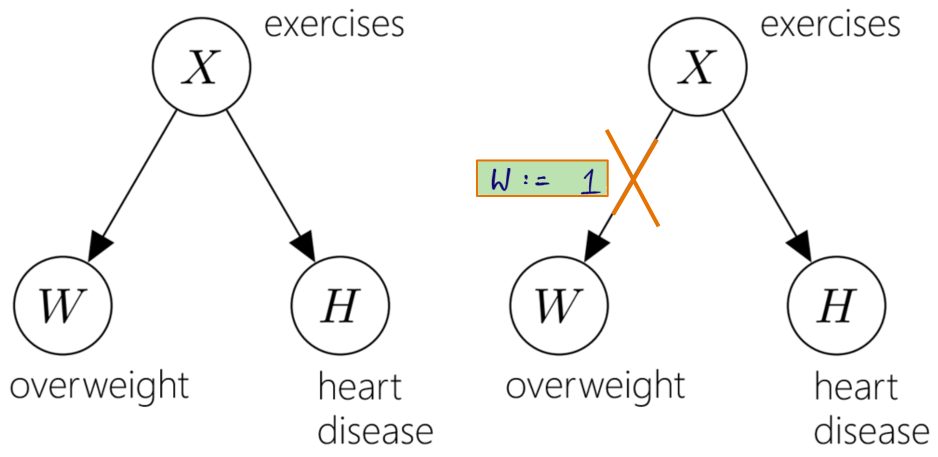
$p(H = 1|W = 1) = \frac{1}{3}$

- Thus $p(H = 1|W = 1) > p(H = 1)$. Does this mean W cause H? NO
- If it did, then intervening/acting on W would change H. Lets try it.

A "program" as a Structural Equation Model (SEM)

- Each of the two programs we saw actually define an SEM.
- Each comes with an acyclic assignment graph called a *causal graph*.

- One variable causes another if there exists a directed path between the two.
- From the graph (also from the program) we see that X *causes* each of W *and* H.
- Causes are your ancestors (direct or indirect causes).



1. Sample Bernoulli random vars
   $U_1 \sim B(1/2)$, $U_2 \sim B(1/3)$, $U_3 \sim B(1/3)$
2. $X := U_1$ (exercise)

   $\boxed{3.\ W := 1}$

3. $W :=$ if $X=1$ then 0 else U_2 (overweight)
4. $H :=$ if $X=1$ then 0 else U_3 (heart disease)

Formally: Structural Equation Model (SEM)

SEMs consist of:

- A list of assignments to generate a distribution on $(X_1, \dots, X_m)$ from independent random (noise) variables, $(N_1, \dots N_{m'})$.
- Must be acyclic assignments (graphical models need not be).

Example: N, N' indep. noise

$X := N$

$Z := 2X + N'$

$Y := (X + Z)^2$

model $\boldsymbol{M}$

Example: N, N' indep. noise

$X := N$

$\cancel{Z := 2X + N'}$ $\boxed{Z = 8}$

$Y := (X + Z)^2$

model $\boldsymbol{M}[Z := 8]$

"probability of event after applying do operator": $\mathbb{P}\{E \mid \mathrm{do}(X := x)\} = \mathbb{P}_{M[X:=x]}(E)$

Causal effects

- Often X denotes the presence or absence of an intervention or treatment.
- $p(Y = y|do(X = x))$ is called the causal effect of X on Y.
- The *average treatment effect* is in turn given by $E[Y = y|do(X = 1)] - E[Y = y|do(X = 0)]$.
- It tells us how much treatment (causally) increases the expectation of Y relative to no treatment (action $X := \mathbf{0}$ vs $X \coloneqq 1$).

A fundamental question in causality

$\boxed{\text{When/how can we estimate causal effects from observational data?}}$

$$E[Y = y|do(X = 1)] - E[Y = y|do(X = 0)]$$

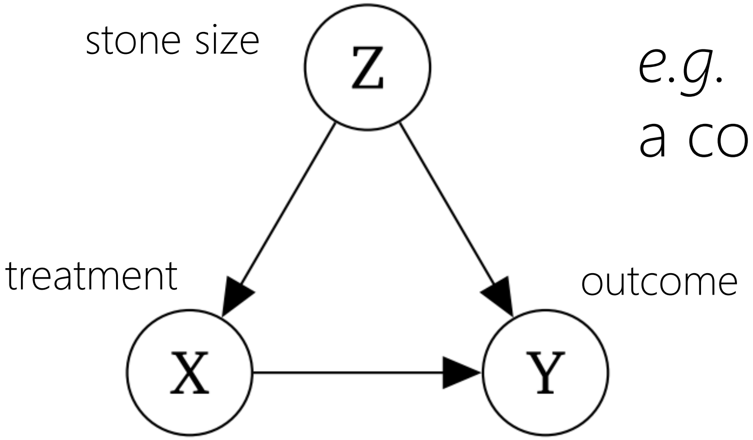
Equivalently:

When/how can we express do-interventions (actions) with a formula that involves only conditional probabilities?

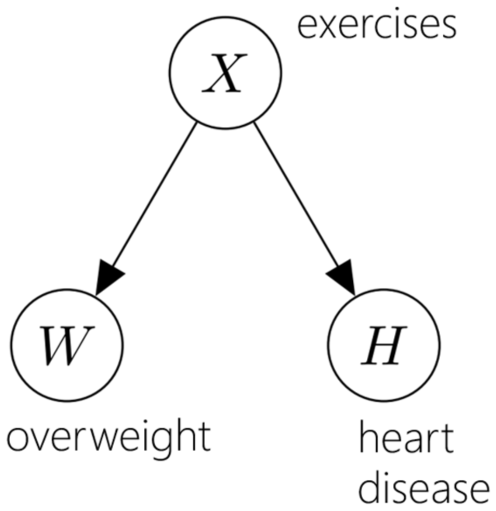
$$p(Y = y|X = x) \neq p(Y = y|do(X = x))$$

Problem of confounding: doing vs observing

Two variables, X and Y are *confounded* if in a causal graph some *confounding variable*, Z, is pointing (causally effecting) each of X and Y:



e.g. "stone size" is a confounder



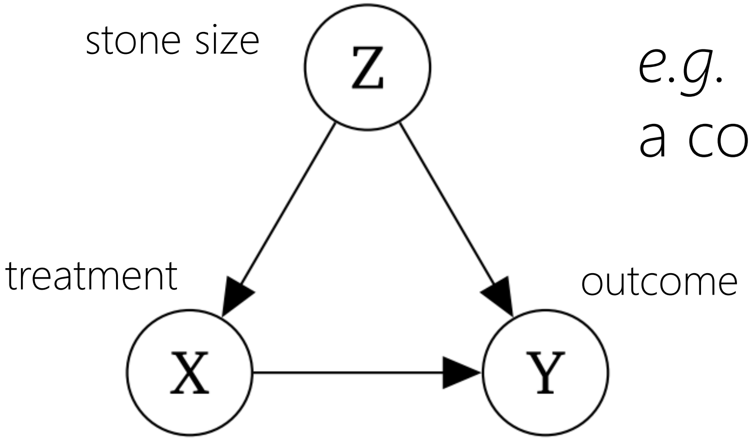
e.g. "exercise" is a confounder

In such a scenario, $p(Y = y|X = x) \neq p(Y = y|do(X = x))$

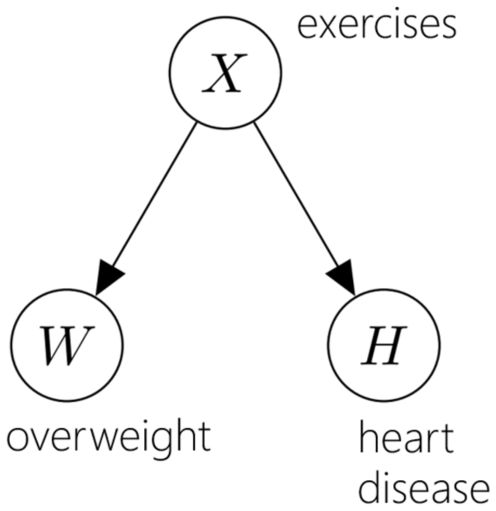
$p(H = h|W = w) \neq p(H = h|do(W = w))$

Problem of confounding: doing vs observing

So how to estimate $E[Y = y|do(X = 1)] - E[Y = y|do(X = 0)]$ with only observational data?



e.g. "stone size" is a confounder



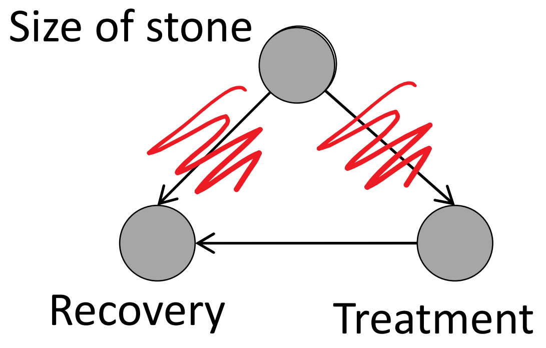
e.g. "exercise" is a confounder

In such a scenario, $p(Y = y|X = x) \neq p(Y = y|do(X = x))$

$p(H = h|W = w) \neq p(H = h|do(W = w))$

Eliminate confounding from observational analysis

- To eliminate confounding, we need to hold the confounding variable constant in our analyses (called *controlling* for that variable).
- To control for kidney stone size, we must compute the treatment effect for each group (stone size) separately.
- Then we can average the effects from each group to get the overall effect.



To do so, we use the *adjustment formula:*

$$\mathbb{P}(Y = y \mid \mathbf{do}(X := x) = \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z).$$

Then we can easily compute the treatment effect:
$E[Y = y|do(X = 1)] - E[Y = y|do(X = 0)]$!

Critically, this requires knowledge of…?

…of the SEM to read off the confounding variables!

Side note on the adjustment formula

The *adjustment formula:*

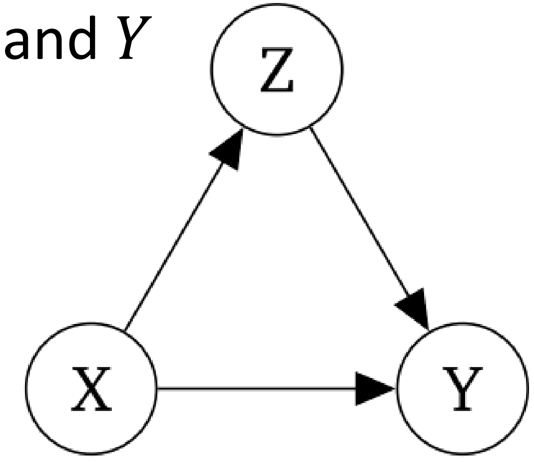
$$\mathbb{P}(Y = y \mid \mathbf{do}(X := x)) = \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z).$$

In contrast to the law of total probability:

$$P(Y|X) = \sum_z P(Y, Z|X) = \sum_z P(Y|X, Z)P(Z|X)$$

Eliminating confounding

Z is a mediator between X and Y



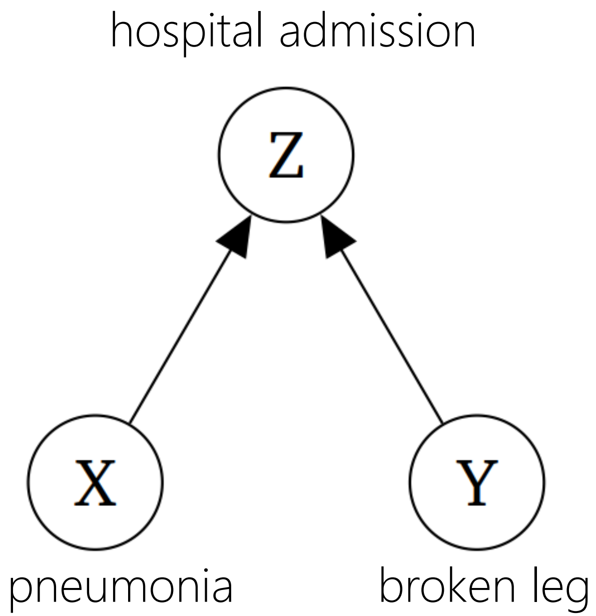
- Should we control for as many variables as we can get our hands on?
- No: we should not control for *mediators* (variables on a "direct path" between X *and* Y).
- Controlling for mediators will *reduce* the effect size we find between X *and* Y.
- e.g. if control for tar deposit, then will reduce the ability to see causal effect $X \rightarrow Y$.

“Collider” variables

- Collider variables are those with incoming effects from X and Y.
- Conditioning on colliders can create anti-correlation between X and Y when they are actually uncorrelated in the population (“Berkson’s law” or “collider bias”).

e.g.

- If we are in the hospital and observe that an individual has a broken leg, what does that tell us about the patient having pneumonia?
- Since a broken leg is a sufficient cause for being in the hospital, it “explains away” the other causes:
- If we condition on Z we might incorrectly conclude that X and Y are anti-correlated.



Z is a collider between X and Y

What have we bought ourselves?

1. We have an intuition for how confounding variables can mess up observational analyses (e.g. stones, treatment & Simpson's paradox).
2. We understand how *doing* is different from *conditioning:*
3. When two variables are confounded, conditioning is the not the same as doing.
4. Introduced the bare bones concepts of SEMs, and how they let us reason about confounding and perform control/adjustment.
5. But all of this formalism is <u>only as good as the SEM is an accurate depiction of the true causal mechanisms</u>!
6. How might we create an accurate SEM?
7. How do we know if our SEM is the correct causal model?

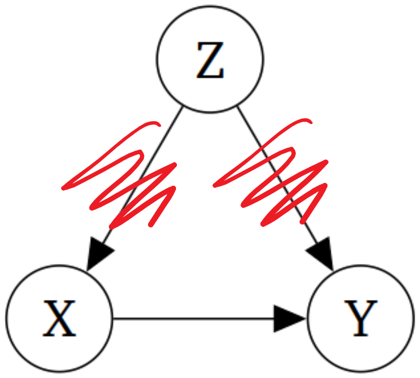
Determining validity of causal models

- In mainstream (non-causal) ML, we can estimate how good a model is using cross-validation.
- There is no analog for determining the validity of an SEM.
- We must use domain knowledge, expertise, or RCTs.
- To get a causal effects from observational studies we need to make assumptions about the "causal story" formally using causal models/graphs, which encode our assumptions about the world.
- Given a causal graph, we can decide what variables are confounders, and the do the appropriate computations.

EXTRA SLIDES

Proof of the Adjustment Formula (simple case)

$$\mathbb{P}(Y = y \mid \mathbf{do}(X := x)) = \sum_z \mathbb{P}(Y = y \mid X = x, Z = z)\mathbb{P}(Z = z).$$



Proof of Adjustment Formula. First, note that

$$\mathbb{P}(Y = y \mid \mathbf{do}(X := x),\ Z = z) = \mathbb{P}(Y = y \mid X = x,\ Z = z)$$

since fixing the value of Z blocks the confounding influence of Z in the causal graph (Figure 14.1). Then, by applying the law of total probability to the model where we make the do-intervention $\mathbf{do}(X := x)$,

$$\begin{aligned}\mathbb{P}(Y = y \mid \mathbf{do}(X := x)) &= \sum_z P(Y = y \mid \mathbf{do}(X := x),\ Z = z)\mathbb{P}(Z = z) \\ &= \sum_z P(Y = y \mid X = x,\ Z = z)P(Z = z).\end{aligned}$$