Causal Inference for ML Practitioners

Colin Gray
Senior Data Scientist @ Netflix
November 2024

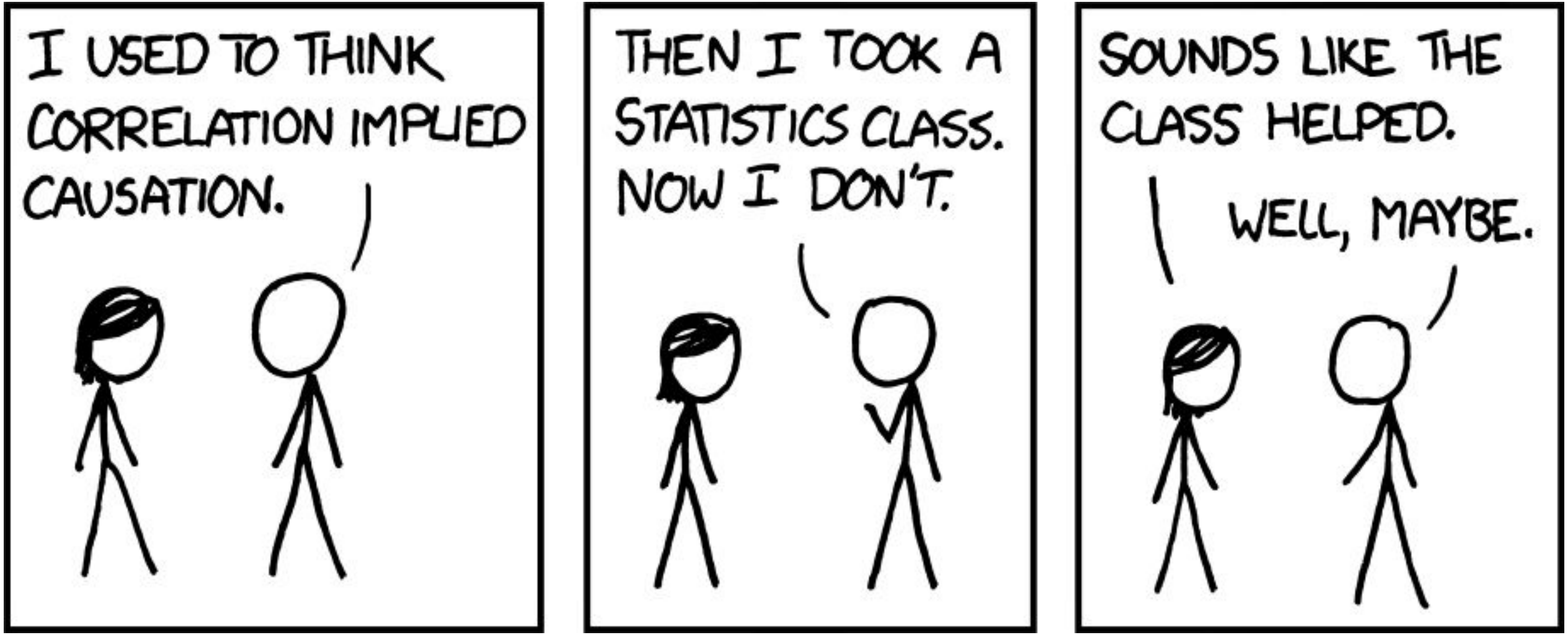
My Goal Today

- Data scientists often know some causal inference concepts, but don't have a clear sense of how they fit together.

- This lecture is meant to give you <u>*intuition*</u> and <u>*structure*</u>
  - We won't cover all of causal inference in one lecture…
  - …but if you hear a common causal inference term, you should know what problem it's addressing & where to learn more!

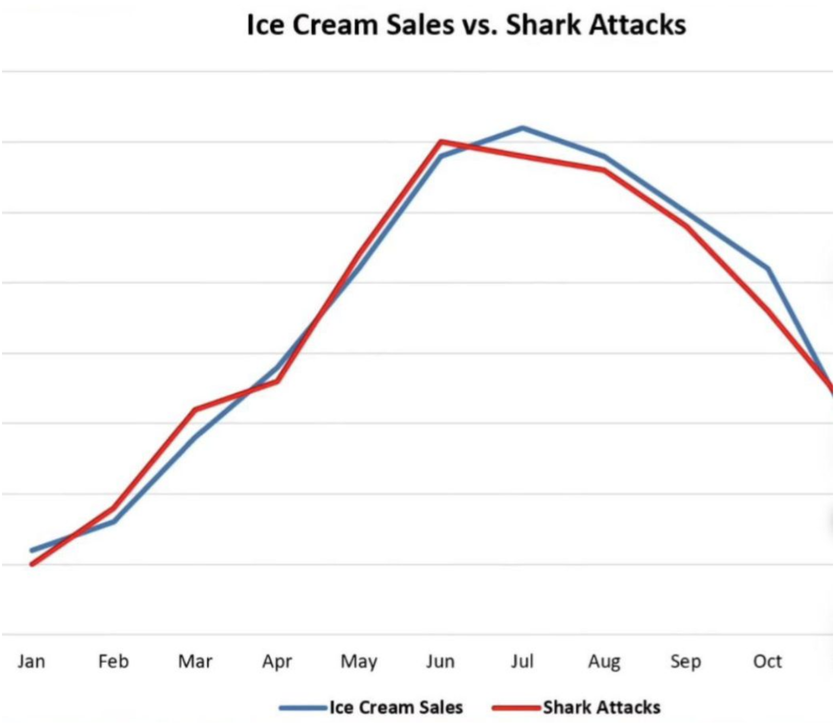
Outline

1. **Correlation != Causation**
2. Adjusting for Confounders
3. Formal Framework
4. Integrating ML

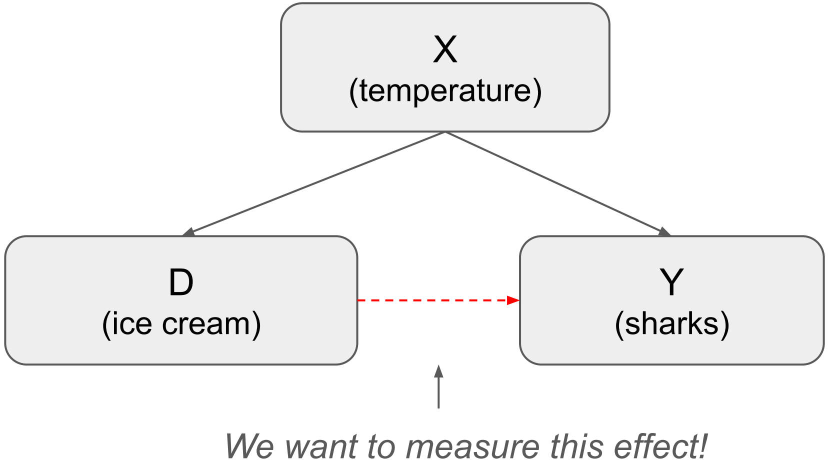


https://xkcd.com/552/

An Extreme Example

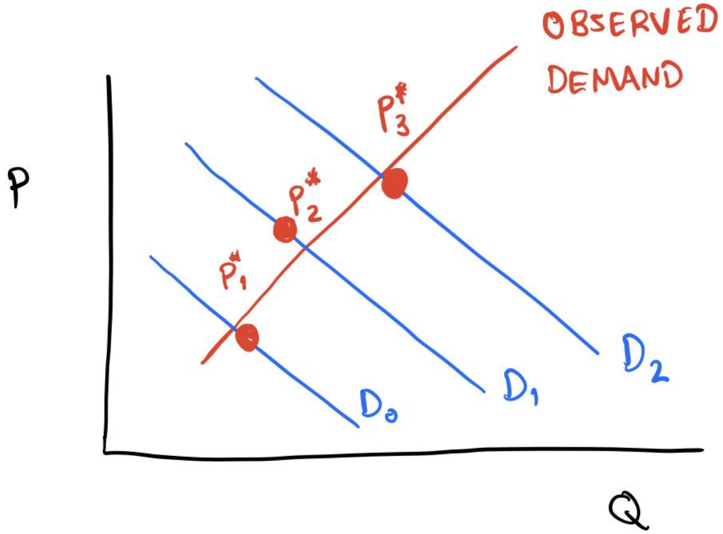


https://www.reddit.com/r/SpuriousCorrelations/comments/12kt1gj/shark_attack_vs_ice_cream_sales/



A Less Extreme Example

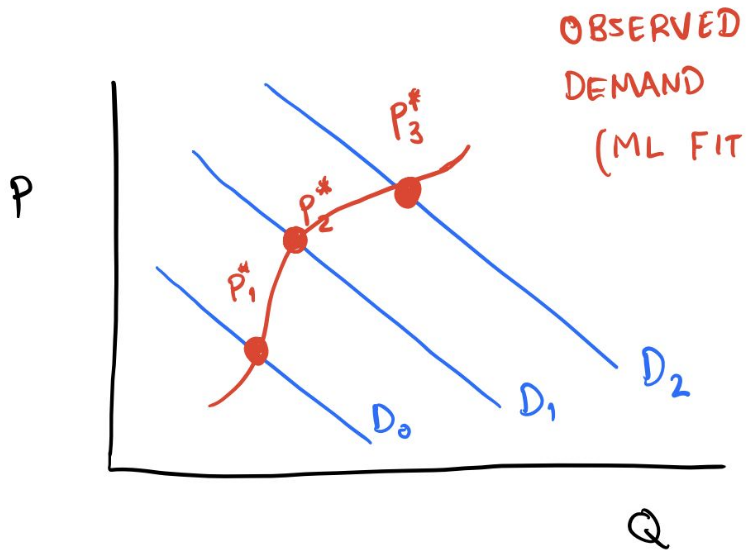
- In raw data, we sometimes see upward-sloping demand curves
  - High prices -> more sales?
  - No! Usually indicates increasing popularity (but data doesn't know that)



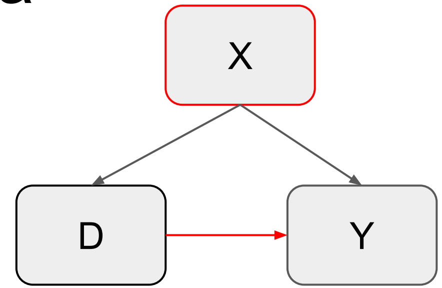
Machine Learning (Alone) Doesn't Fix This

- ML is *great* at model fit…
    - Predict the next word in a sequence
    - Predict orders in specific region
    - Predict likelihood that a customer clicks

- …but model fit is not the issue here!
    - Fitting patterns doesn't reveal causal drivers



Solution #1: Adjust Using Existing Data



- At each <u>*fixed*</u> X, do we see a relationship btw D and Y?
  - Section 2 will cover *how* to do these adjustments

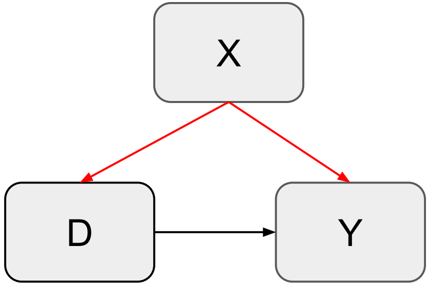
- Pro: Uses existing data (cheap)

- Con: Requires assumptions
  - Assume D→Y, *not* Y←D
  - Assume we know X

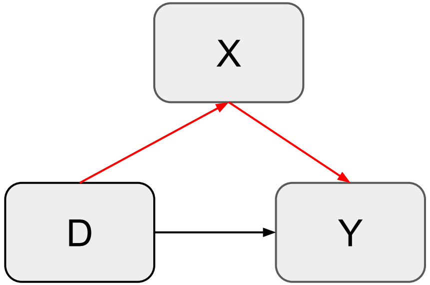
Warning: We don't want to adjust for everything!

- *Do* Adjust for **Confounders** (Today's Focus)

D = ice cream
X = temperature
Y = shark attacks



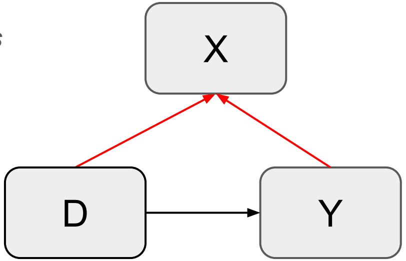
- *Do Not* Adjust for **Mediators**



D = education
X = profession
Y = income

- *Do Not* Adjust for **Colliders**

D = attractiveness
X = movie roles
Y = acting talent

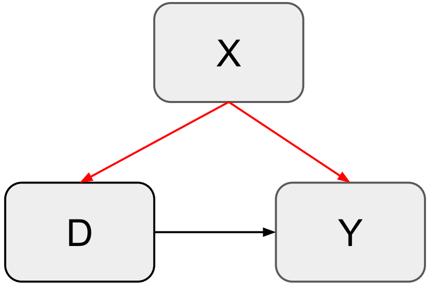


- We'll stick with confounders today, but always think through your DAG…

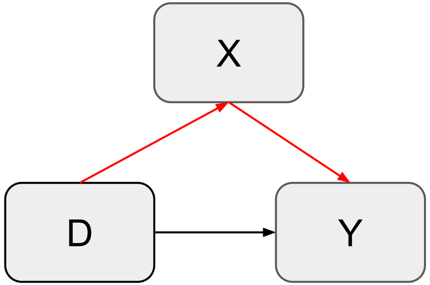
Warning: We don’t want to adjust for everything!

- *Do* Adjust for **Confounders** (Today’s Focus)

D = smoking
X = drinking
Y = cancer



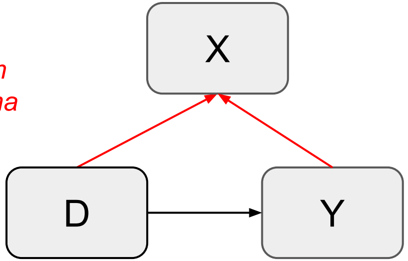
- *Do Not* Adjust for **Mediators**



D = exercise
X = weight loss
Y = heart disease

- *Do Not* Adjust for **Colliders**

D = pollution
X = hospitalization
Y = chronic asthma



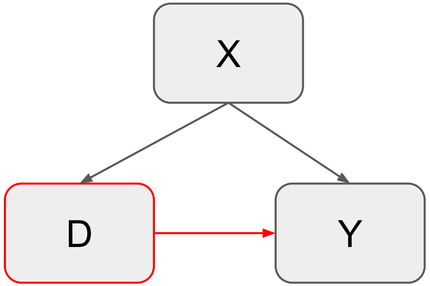
- We’ll stick with confounders today, but always think through your DAG…

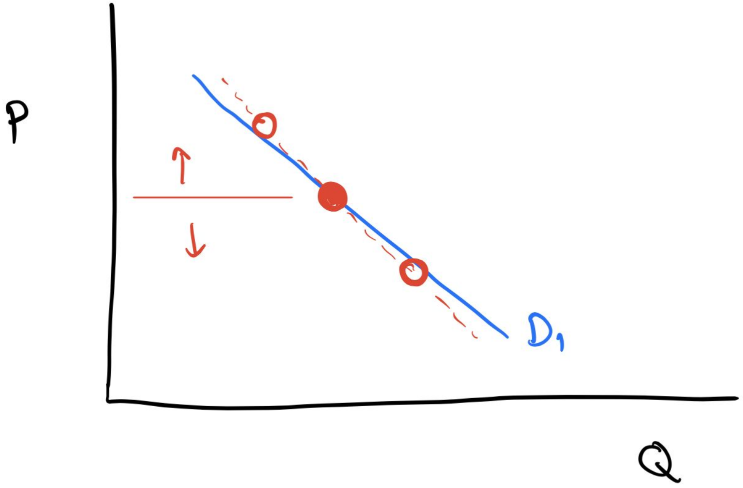
Solution #2: Experiment

- Randomly alter the variable of interest
  - Give people free ice cream.
    - Do shark attacks increase?
  - Price +/- 5%
    - Do sales respond?

- Pro: Minimal assumptions
  - Don't have to specify D→Y or Y←D
  - Don't have to specify X

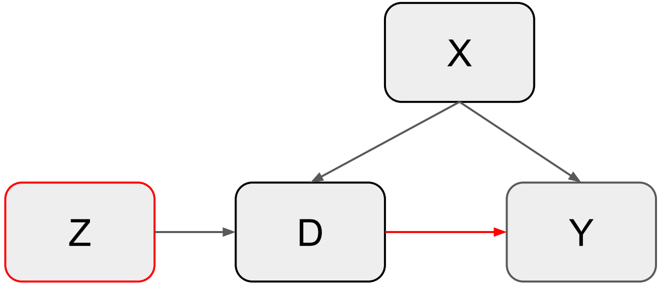
- Con: Uses bespoke data (expensive)





Solution #3: Quasi-Experiments

- If we have some external variable that *only* impacts D, then can we mimic an experiment?
  - Z is an "instrumental variable"



- Pro: Uses existing data (cheap), few assumption on X

- Con: Requires separate assumptions
  - Are we confident that Z affects D, and nothing else?
  - *Don't try this until you have a lot of experience… But you should know it exists!*

Recap: Correlation != Causation

- Often, raw comparisons do *not* capture underlying causal impacts
  - We can't just "ML" our way out of the problem
- We might be able to adjust existing data for confounders
  - Very useful bag of tricks (coming up), but convince yourself that you're actually dealing with *confounders* (not *mediators* or *colliders*)
- Experiments sidestep this, but they're expensive
  - Advanced techniques try to mimic experiments, but don't try at home (yet)

Outline

1. Correlation != Causation
2. **Adjusting for Confounders**
3. Formal Framework
4. Integrating ML

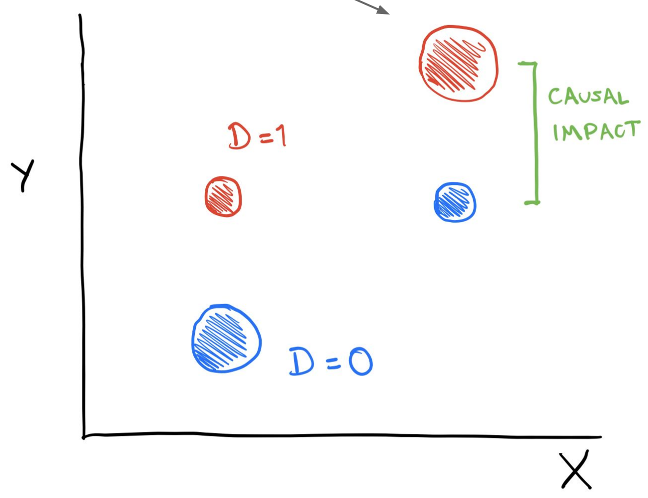
A Thought Experiment

- Let’s say we *knew* there was a single confounder X
  - This is a good candidate for Solution #1 in the previous section
  - Let’s also imagine our data isn’t noisy (for now), just for illustration

- How would we adjust for this in our (non-experimental) data?
  - We have a handful of approaches, which might ring some bells

Bigger dots contain more people (N)

Set Up



- Each person (i) has outcome (Y), treatment (D), confounder (X), noise (e)
- "True" model: $Y_i = \alpha + \beta D_i + \gamma X_i + e_i$
- Will comparing treatment vs control averages yield the causal effect ("beta")?
- Our example says "no"!
  - True impact is 1
  - E[Y | D=1] - E[Y | D=0] = 1.75 - 0.25 = 1.5 > 1
  - The problem? X associated with D *and* Y!

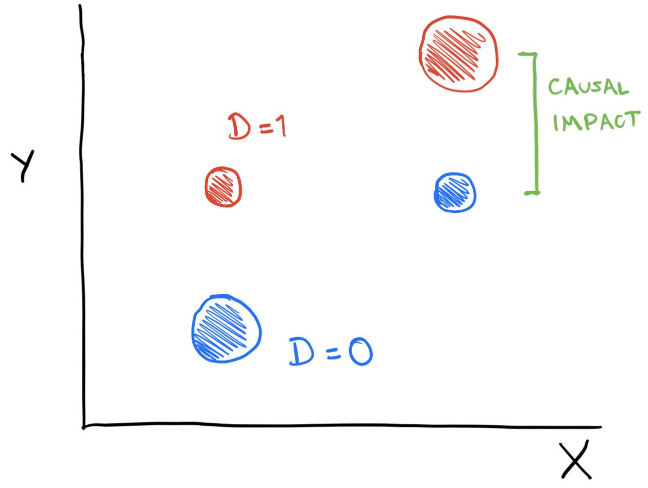
D	X	N	Y
0	0	3	0
0	1	1	1
1	0	1	1
1	1	3	2

Solution #1: “Match” on X

- Compute difference within each value of X, then aggregate

- Let’s try this in our example
  - E[Y|D=1, X=0] - E[Y|D=0, X=0] = 1 - 0 = 1
  - E[Y|D=1, X=1] - E[Y|D=0, X=1] = 2 - 1 = 1
  - Average Difference = 1
  - It worked!

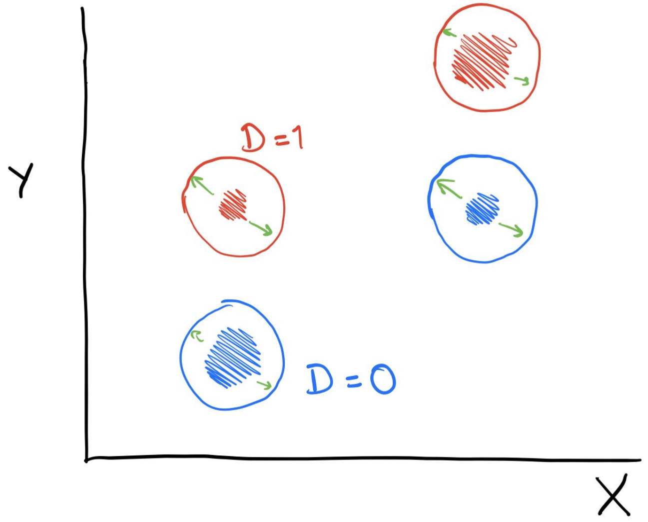
- <u>Catch</u>: Tricky if X has continuous and/or many dimensions
  - “Curse of dimensionality”



D	X	N	Y
0	0	3	0
0	1	1	1
1	0	1	1
1	1	3	2

Solution #2: Reweight the Data

- Eliminates relationship btw D and X
  - E[D|X] is called the "propensity score", denoted p(X)
  - Weight "surprising" observations more
    - 1 / p(X) if D = 1
    - 1 / (1-p(X)) if D = 0
  - In practice, often have to *estimate* the propensity score



- Let's try this in our example
  - E[D|X=0] = p(0) = 0.25
  - E[D|X=1] = p(1) = 0.75
  - Difference (Weighted) Means = 1.5 - 0.5 = 1
  - It worked!

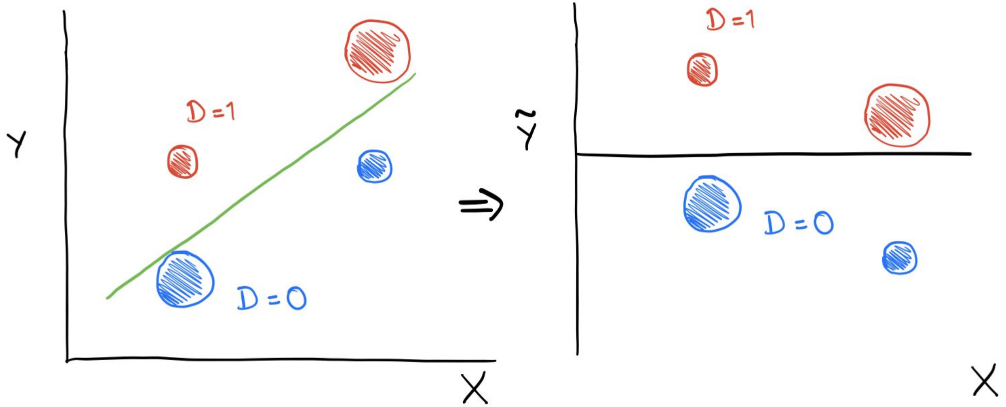
- <u>Catch</u>: Unstable if p(X) close to 0 or 1

D	X	N	Y	Divisor	Weighted N
0	0	3	0	1-P = 0.75	4
0	1	1	1	1-P = 0.25	4
1	0	1	1	P = 0.25	4
1	1	3	2	P = 0.75	4

Solution #3: “Tilt” the Data

- A reasonable first idea:
    - Find “unexpected” part of Y
    - $\tilde{Y}_i \equiv Y_i - E[Y_i|X_i]$
    - Difference using that value instead…

- This doesn’t *quite* work…
    - E[Y|X=0] = 0.25
    - E[Y|X=1] = 1.75
    - Difference (Adjusted) Means
        - 0.375 - (-0.375) = 0.75 < 1

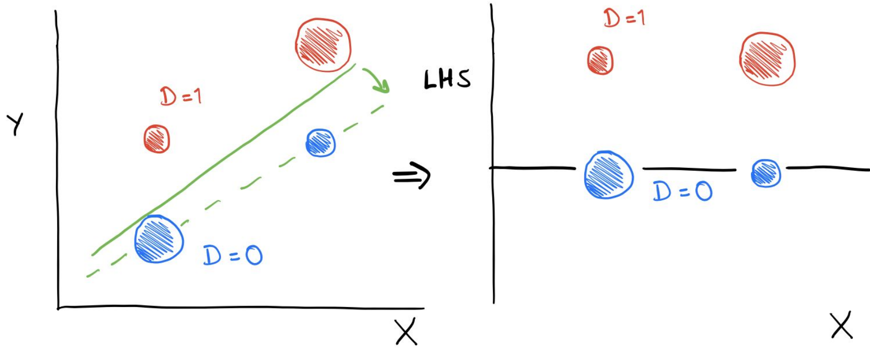


D	X	N	Y	E[Y | X]	Y_tilde
0	0	3	0	0.25	-0.25
0	1	1	1	1.75	-0.75
1	0	1	1	0.25	0.75
1	1	3	2	1.75	0.25

Solution #3: "Tilt" the Data

- Better idea…
  - Find "unexpected" Y and "unexpected" D

$$Y_i - E[Y_i|X_i] = \beta(D_i - E[D_i|X_i]) + e_i$$
$$\Rightarrow Y_i - E[Y_i|X_i] + \beta E[D_i|X_i] = \beta D_i + e_i$$



- This works!
  - Satisfied under correct beta of 1
  - (How do we identify correct beta? Next slide.)

- Catch: Less transparent weightings when effects vary across X

D	X	N	Y	E[Y | X]	E [D | X]	LHS | beta=1
0	0	3	0	0.25	0.25	0
0	1	1	1	1.75	0.75	0
1	0	1	1	0.25	0.25	1
1	1	3	2	1.75	0.75	1

Solution #3: Multivariate Regression in Disguise!

- To operationalize #3, we don't have to subtract out these expected values…

$$Y_i - E[Y_i|X_i] = \beta(D_i - E[D_i|X_i]) + e_i$$

- *If E[.|X] are linear in X*, then we get numerically equivalent numbers by controlling for X in a multiple regression!

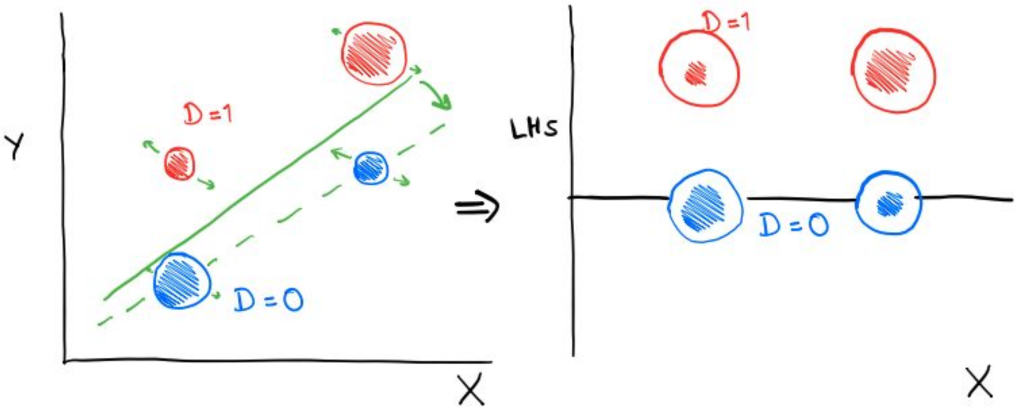
(Adjusted) Treatment Effect

$$Y_i = \alpha + \beta D_i + \gamma X_i + u_i$$

  - Frisch-Waugh-Lovell Theorem (FWL)
  - *This* is (part of) why people like regression so much!

More Exotic: Mix & Match

- "Doubly-Robust" models
  - e.g. reweight, then regress
  - e.g. match, then regress
  - Our estimate will be correct if *either* model is correct



D	X	N	Y	Weighted N	E[Y | X]	E [D | X]	LHS
0	0	3	0	4	0.25	0.25	0
0	1	1	1	4	1.75	0.75	0
1	0	1	1	4	0.25	0.25	1
1	1	3	2	4	1.75	0.75	1

Recap: Adjusting for Confounders

- Often, raw comparisons do *not* capture underlying causal impacts due to *confounders*

- If we know the confounders, we have multiple ways to “undo” their effects
  - “Match” on X
  - “Reweight” (or match) on the propensity score
  - “Tilt” the data, operationalized via multivariate regression
  - Some more exotic variants (e.g. doubly-robust models)…

- We haven’t discussed <u>*what*</u> to adjust for
  - What if X impacts Y in complex ways (e.g. nonlinear)?
  - What if we don’t even know which X variables matter?
  - We’ll come back to this…

Outline

1. Correlation != Causation
2. Adjusting for Confounders
3. **Formal Framework**
4. Integrating ML

Benefits of Formality

- Hopefully you have some intuition from graphs & examples
  - Let's solidify our understanding by showing it in math

- This will also let us repurpose some familiar statistical tools
  - e.g. it would be *really* useful to have some measurements of uncertainty

Potential Outcomes

- Let's posit that two *potential outcomes* simultaneously exist
  - In one state of the world, person *i* is not treated
  - In the other, person *i* is treated
  - (Does this make sense? Ask the philosophers… but it's very useful.)

$$Y_i(0) = \alpha + \gamma X_i + e_i$$
$$Y_i(1) = \alpha + \beta D_i + \gamma X_i + e_i$$

- We (the scientist) only get to see one potential outcome per person
  - The "fundamental problem of causal inference"

What We Want != What We Have

- Our *__estimand__* (or "target parameter") is the difference we *would* see if we could observe both potential outcomes

$$E[Y_i(1) - Y_i(0)] = E[(\alpha + \beta + \gamma X_i + e_i) - (\alpha + \gamma X_i + e_i)] = \beta$$

- But we're stuck with our *__estimator__* (observed averages) instead…

$$\begin{gathered} E[Y_i|D_i = 1] - E[Y_i|D_i = 0] \\ = (\alpha + \beta + \gamma E[X_i|D_i = 1]) - (\alpha + \gamma E[X_i|D_i = 0]) \\ = \beta + \gamma(E[X_i|D_i = 1] - E[X_i|D_i = 0]) \end{gathered}$$

Reinforcing Our Intuition

- Look closer… here are the principles we developed in Section 2!
  - This is sometimes called "omitted variable bias" (OVB)

"Estimator" we have

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$
$$= (\alpha + \beta + \gamma E[X_i|D_i = 1]) - (\alpha + \gamma E[X_i|D_i = 0])$$
$$= \beta + \gamma(E[X_i|D_i = 1] - E[X_i|D_i = 0])$$

True Effect
("Estimand" we want)

X impacts Y

X impacts D

Match or pscore weight to kill (.) term

Use regression to eliminate this relationship

After Adjusting for Confounders

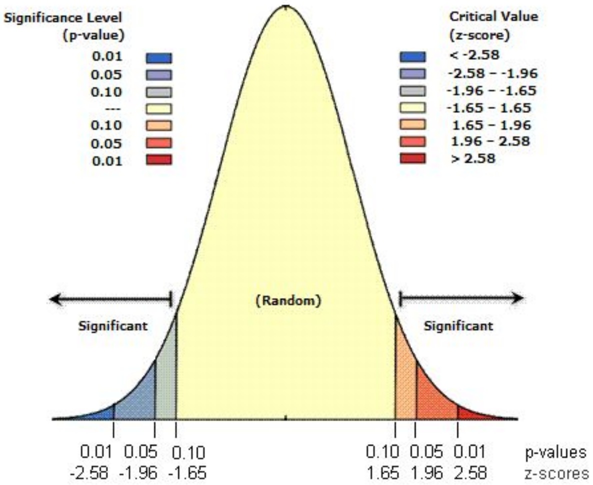
- Suppose we already did our preferred confounder adjustment
  - e.g. matching, pscore reweighting, regression adjustment

- Our estimated effect is back on track, and is the difference approximates two (adjusted) averages

$$\hat{\beta} \rightarrow E[Y_i(1)] - E[Y_i(0)]$$

Quantifying Uncertainty

- We know how sample averages behave (in large samples)!
  - Thanks to the Central Limit Theorem they are approximately Gaussian
  - Difference between two Gaussian random variables is also Gaussian
  - *So* we can use familiar t-tests and z-scores to measure uncertainty, even after adjustments!



https://desktop.arcgis.com/en/arcmap/latest/tools/spatial-statistics-toolbox/what-is-a-z-score-what-is-a-p-value.htm

Operationalizing

- If you're *matching* or *pscore reweighting*, then you can either:
  - (1) Use standard t-tests (on matched/weighted data)
  - (2) Get a z-score by running a simple regression (on matched/weighted data)

$$Y_i = \alpha + \beta D_i + e_i$$

Control Group Average

Treatment Effect
w/ Standard Errors

- If you're adjusting for confounders in a *regression*, you can simply add your X variables
  - Thanks to FWL Theorem

$$Y_i = \alpha + \beta D_i + \gamma X_i + u_i$$

(Adjusted) Treatment Effect
w/ Standard Errors

Example

```python
# Import packages
import numpy as np
import pandas as pd
import statsmodels.api as sm

# Generate fake data
N=1000
X = np.random.uniform(size=N)
D = (X + np.random.uniform(size=N) > 1).astype(int)
e = np.random.normal(size=N)/4
Y = 1 + D + X + e
df = pd.DataFrame({'Y': Y, 'D': D, 'X': X})

# Models
unadjusted = sm.OLS.from_formula('Y ~ 1 + D', data=df).fit(cov_type='HC1')
print(f"\n UNADJUSTED: \n {unadjusted.summary()}")

adjusted = sm.OLS.from_formula('Y ~ 1 + D + X', data=df).fit(cov_type='HC1')
print(f"\n REGRESSION ADJUSTED: \n {adjusted.summary()}")
```

True effect = 1

```
UNADJUSTED:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.777
Model:                            OLS   Adj. R-squared:                  0.777
Method:                 Least Squares   F-statistic:                     3512.
Date:                Thu, 29 Aug 2024   Prob (F-statistic):               0.00
Time:                        10:03:49   Log-Likelihood:                -359.57
No. Observations:                1000   AIC:                             723.1
Df Residuals:                     998   BIC:                             733.0
Df Model:                           1
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.3405      0.016     86.069      0.000       1.310       1.371
D              1.2981      0.022     59.261      0.000       1.255       1.341
==============================================================================
Omnibus:                        3.874   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.144   Jarque-Bera (JB):                3.894
Skew:                          -0.132   Prob(JB):                        0.143
Kurtosis:                       2.844   Cond. No.                         2.71
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Not that close…

```
REGRESSION ADJUSTED:
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.885
Model:                            OLS   Adj. R-squared:                  0.885
Method:                 Least Squares   F-statistic:                     3645.
Date:                Thu, 29 Aug 2024   Prob (F-statistic):               0.00
Time:                        10:03:49   Log-Likelihood:                -29.181
No. Observations:                1000   AIC:                             64.36
Df Residuals:                     997   BIC:                             79.09
Df Model:                           2
Covariance Type:                  HC1
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      0.9847      0.017     57.388      0.000       0.951       1.018
D              1.0200      0.018     58.238      0.000       0.986       1.054
X              0.9882      0.032     31.242      0.000       0.926       1.050
==============================================================================
Omnibus:                        1.412   Durbin-Watson:                   1.929
Prob(Omnibus):                  0.494   Jarque-Bera (JB):                1.440
Skew:                           0.049   Prob(JB):                        0.487
Kurtosis:                       2.842   Cond. No.                         5.80
==============================================================================

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)
```

Quite close!

Confidence interval covers true value

Recap

- We posit the existence of “potential outcomes”
  - Gives us a clear “estimand” that we’re targeting
  - The “estimator” we have may or may not align with that target
  - This distinction (what we want vs what we have) is *critical* in causal inference!

- This framework reinforces our intuition from Section 2
  - “Omitted variable bias” puts daylight between our estimand vs our estimator
  - Our various corrections try to re-align them

- This framework puts us back in a familiar statistical sampling paradigm
  - Averaging samples of (potential) outcomes lets us use familiar CLT theorems to quantify uncertainty

Outline

1. Correlation != Causation
2. Adjusting for Confounders
3. Formal Framework
4. **Integrating ML**

What if we don’t know X?

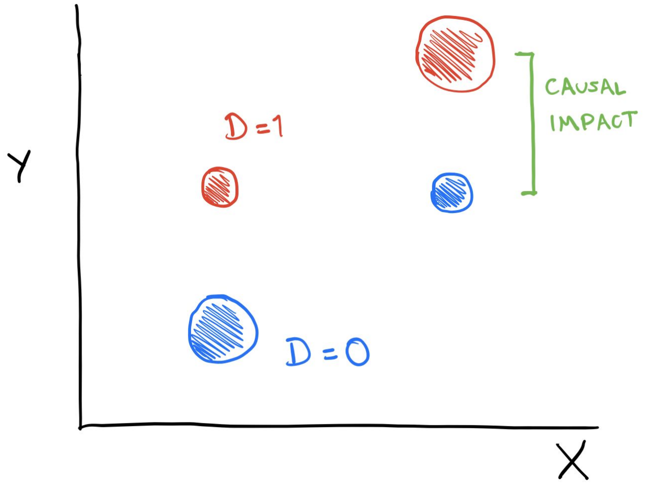
- Up until now, we’ve assumed we know X *and* its functional form (e.g. “linear”)
  - This isn’t a very common situation out in the wild

- “Causal ML” techniques are an increasingly popular way to get a handle on this problem
  - I’ll give you a flavor of how a couple popular techniques work
  - Generally fall under the header of “causal meta-learners”

T-Learners

- Directly predict E[Y|D=0, X] and E[Y|D=1, X] using ML techniques, then compare them

- This works okay in practice
  - Standard errors may be less reliable
  - Can result in overly complex treatment effects, since prediction models both have independent swings
  - Some modifications (e.g. X-Learner) can help



Double Machine Learning (DML)

- Perhaps the most popular Causal ML technique right now

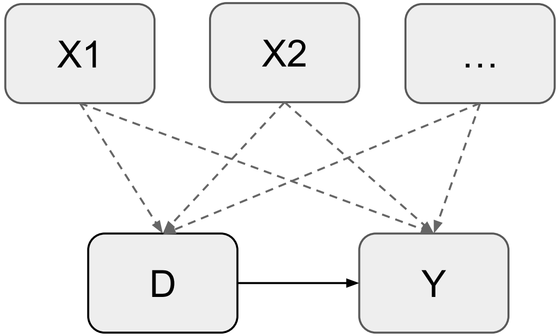
- Based off of the expression we mentioned in Section 2:

$$Y_i - E[Y_i|X_i] = \beta(D_i - E[D_i|X_i]) + e_i$$

- Big Insight: If we have lots of candidate Xs, can we use ML to *predict* E[Y|X] and E[D|X], then plug those in directly?
  - Yes! But only under certain conditions (e.g. good predictors, sample-splitting)
  - Extensions to model how effects vary across individuals (e.g. R-Learner)

Words of Caution

- These ML techniques are *not magic*
    - In some cases, help us sort through candidate X variables & functional forms
    - They do ***not*** help us prove causality, avoid mediators/colliders, or ensure we've caught all confounders
    - Read up, take a course, use with care…



Recap: Integrating ML

- Techniques using ML to address causal questions are increasingly popular
  - You should recognize their names & know how they fit into our framework

- They do ***not*** negate the need for careful causal thinking
  - Typically used to sort through candidate confounders and/or model varying effects
  - The rest of our core causal concerns are alive & well
    - What direction does causality run?
    - Have we captured all confounders?
    - Are we accidentally adjusting for colliders or mediators?
    - Do we have enough data to get a signal?
    - etc…

Outline

Sharks (probably) don’t like ice cream

1. Correlation != Causation
2. Adjusting for Confounders
3. Formal Framework
4. Integrating ML

Matching, pscore-weighting, regressions, etc

Sampling potential outcomes

Helps in some tasks, not magic

Resources to Learn More

- Causal Inference for the Brave and True (Facure)
- Mastering ‘Metrics (Angrist & Pischke)
- Business Data Science (Taddy)
- EconML Documentation (Microsoft)
- Causal ML Book (Chernezhukov et al)